Predicting Antigen Evolution

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Albina Rahim

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Evolution does not happen at random and one strain of a bacterium cannot evolve directly to another arbitrary strain. Nature allows only certain paths in going from one strain to another. Based on this idea, this research aims to identify what the fHbp (factor H binding protein) antigen of the *Neisseria meningitidis* bacterium is likely to mutate into so that vaccines and therapies can be developed in advance for the most likely mutants. *Neisseria meningitidis* is the bacterium that causes the potentially devastating disease meningococcal meningitis in humans.

The target of this research was to generate valid new variants of fHbp based on the characteristics of the existing fHbp sequences. The characteristics were that the sequences had specific invariant regions which flanked restricted variable regions, the mutations in each position of the variable regions were highly constrained, there were peptides in the homologous sequences which were correlated or were co-occurring with each other, and there were regions which were more or less likely to mutate than by chance. As part of determining the characteristics, tertiary structures of the existing sequences were also predicted and energy values of those structures were determined. A pipeline of programs was written to generate variants of the fHbp sequence which satisfied these characteristics. The new variants were studied, their tertiary structures were predicted, and energy values of those structures were determined, similar to what was done for the existing variants. New variants whose associated energy values fell into the range defined by existing sequences were deemed to be "allowable" by nature.

Unfortunately, all of the variants of fHbp generated were valid according to our energy criterion. All generated variant being "allowed" is an unlikely result. Therefore, we conclude that a more stringent methodology for evaluating the viability of fHbp variants is necessary.

Another contribution of our work is the program for generating variants. Like the methodology, it is highly modular, and can be easily used as starting platform for research into additional filtering criteria.

Acknowledgements

I would like to humbly thank:

My supervisor, Dr. Anthony Kusalik for providing me with this opportunity, and for his constant support, encouragement, and guidance throughout my studies.

My committee members, Dr. Ian McQuillan and Dr. Nate Osgood for their helpful comments and suggestions, and Dr. Yu Luo for agreeing to be the external examiner.

My fellow lab mates and friends, Brett, Lingling, Teenus, and Tejumoluwa for their friendship and support. Good luck to each of you in your future aspirations.

My parents and siblings, for their unending love, support, and inspiration.

My best friend and husband, Md Rezaul Karim for your unconditional love and support and for always ensuring I smile even when it seemed tough.

Dedicated to the Almighty for constant support, strength, and guidance.

Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

Ab	Antibody
Ag	Antigen
CCP	Complement Control Protein
CD-HIT	Cluster Database at High Identity with Tolerance
DNA	deoxyribonucleic acid
EMBOSS	The European Molecular Biology Open Software Suite
EXPR	scalar expression
fH	factor H
fHbp	factor H binding protein
gna1870	genome-derived *Neisseria* antigen 1870
HMM	hidden Markov models
LP2086	lipoprotein 2086
MEGA4	Molecular Evolutionary Genetics Analysis software version 4.0
NN	neural network
OMP	outer membrane proteins
PDB	Protein Data Bank
RNA	ribonucleic acid
RMSD	root mean square deviation
TCR	T cell antigen receptor
UniProtKB	Universal Protein Resource Knowledgebase
UniRef	UniProt Reference Clusters
WHO	World Health Organization

Chapter 1

Introduction

Neisseria meningitidis is a major human pathogen that causes the disease meningococcal meningitis. It is a gram-negative, aerobic bacterium which infects only humans. Transmission of this bacterium occurs directly by droplets spread from person to person. The most common symptoms of meningitis include fever, headache, nausea, vomiting, skin rashes, and neck stiffness. The disease can also lead to stupor and coma, or leave the infected individuals with irreparable permanent damage like paralysis or amputated limbs.

Factor H binding protein (known as fHbp, GNA1870, or LP2086) is an antigen which is present on the surface of all strains of the *Neisseria meningitidis* bacterium. fHbp binds with factor H (fH), a key regulatory component of the human complement immunological pathway. It enables the meningococcus to avoid innate immune responses by inhibiting complement-mediated lysis in human plasma (Section 2.6). Thus fHbp is important for survival of the *N. meningitidis* in human blood and human sera, and in the presence of antimicrobial peptides. This makes fHbp a potential vaccine antigen [68, 70]. Hence, it was chosen for this research.

The goal of this research is to aid in the development of vaccines and therapies for the meningitis disease based on the concept of predicting the evolution of the fHbp antigen. Evolution does not happen randomly and the transformation of this bacterial antigen is allowed to follow only in certain evolutionary paths which in turn are influenced by genetic diversity of the sequences. In this research we tried to determine what the fHbp antigen will mutate into, with the ultimate goal of generating valid new, likely variants of fHbp. The new variants will then facilitate the development of vaccines and therapies in advance of the appearance of mutant strains containing them.

In order to generate the new variants of fHbp that have not yet appeared but which nature might "allow", it is essential to study the characteristics of the existing fHbp sequences so as to be able to generate new sequences (variants) which have those same characteristics. The relevant characteristics include specific properties of the fHbp sequences, like having restricted variable regions flanked by particular invariant segments, amino acid substitution domains in each position along the sequence alignment, peptides in the sequences which are either correlated (indicating to what extent a mutation in one position of an alignment affects the presence or absence of other amino acids in another position) or are co-occurring (peptides in different positions along an alignment

existing together in a pattern) with each other, sites where the amino acids tend to mutate (positive selection sites) and sites where the amino acids tend to remain conserved (negative selection sites), and the tertiary structures of the sequences and their energy values. A computer program, written as part of this research work, then uses these characteristics to guide the generation of new variants of fHbp. The characteristics of the new variants are then verified with those of the already existing fHbp sequences in order to determine whether they have the same properties. The variants that have the necessary properties are the new fHbp sequences which have not as yet seen but which nature will "allow".

For this research all available fHbp sequences are collected from the UniProtKB database (Section 4.1). After collecting the set and removing duplicate sequences, a multiple sequence alignment is performed on the set, and later investigated. This is done in order to see whether the sequences all satisfy the criteria of starting with a specific amino-terminal repetitive element followed by segments of restricted variable regions flanked by the invariant segments (as is classified by Beernink et al. [34] in their work). The restricted variable regions are then studied to determine the correlation and co-occurrence in the mutation of the amino acids in the different positions of fHbp. After this the constraints of the restricted variable positions are characterized, such as whether one amino acid has higher possibility of substituting another amino acid or if mutation between two amino acids involves presence of an intermediate amino acid, etc. The tertiary structures of the original fHbp sequences are then predicted and their energy values are recorded. This is done in order to determine the valid boundary set by the highest energy value that an existing fHbp sequence can have. Based on all the constraints determined for the fHbp sequences, a computer program is developed which generates new variants of fHbp sequences. The new variants are then examined to determine which of the amino acid substitutions satisfy the constraints determined in the previous steps. Any variants that do not satisfy all of the constraints are discarded. In addition, any duplicates (of variants or of the original set of sequences) are discarded. In the next step, the tertiary structures of the new variants are predicted, after which their energy values are determined. The energy value for each variant is then compared to those for the original fHbp sequences to determine whether it is below the valid energy boundary. It is assumed that molecular variants that pass this test will be "allowed" by nature but have not as yet appeared.

Unfortunately, the 300 new variants generated in this research were all found to be valid, which was an unlikely result. At least some variants whose energy values were not within the valid range were expected due to the stochastic nature of the generation process. Some of the possible reasons behind such results are the absence of bio-chemical verification of these new variants in wet labs and the limited number of structural templates used for predicting the tertiary structures of the fHbp sequences and variants. It could also be that generation of 300 new variants was too few to result in creating diverse variants and more number of variants should be generated.

An important note about this research is that the methodology and the computer program written for generating the new variants are both modular in nature and quite extensible, thereby laying the groundwork for further work on this topic.

1.1 Thesis Organization

The thesis is organized as follows:

- Chapter 2 describes the background knowledge for better comprehension of this research.
- Chapter 3 outlines the goal of this research.
- Chapter 4 discusses the data collected for this research and the methodology followed. Parts of the results are also included in this chapter since some steps of the methodology were dependent on the results generated in the previous steps. This is done for better understanding of the methodology of the work.
- Chapter 5 presents and evaluates the results.
- Chapter 6 discusses the conclusion of this research and possible future work.

Chapter 2

Background Information

To better understand this research some background knowledge is required. This background knowledge relates to protein sequence databases, genome annotations, multiple sequence alignments, antigens, vaccines, the human complement system, *Neisseria meningitidis* and its antigen fHbp, the concept of evolution, and protein structures and their energies. Sections 2.1 to 2.12 give brief descriptions of these topics, aiming to give the reader an introduction to predicting antigen evolution and its associated advantage in vaccine development.

2.1 Protein Sequence Databases and Genome Annotation

A variety of protein sequence databases exist, ranging from simple sequence databases to expertly curated universal databases covering all species. DDBJ, GenPept, NCBI, UniProtKB, and PIR are some examples. Such databases play a significant role as central comprehensive resources of protein information [32].

Genome annotation is the process of attaching biological information to sequences and is done in two steps [7]:

1. identifying elements of the genome, a process called gene prediction, and
2. attaching biological information to these elements.

There are structural annotations consisting of information on gene structure, coding regions, etc. and there are functional annotations consisting of biological information like biochemical function, biological function, etc. [7].

2.2 Multiple Sequence Alignments

Multiple sequence alignments are common operations in bioinformatics and are widely used in all areas of nucleotide and protein sequence analysis. They are needed whenever sets of homologous sequences are compared and are an essential precursor to numerous further analyses. ClustalW, ClustalX, T-Coffee, MUSCLE, MAFT, PROBCONS, Kalign are some of the commonly used multiple sequence alignment tools [79].

2.3 Innate and Adaptive Immunity

Innate immunity forms the first line of defence of the human body to infection and brings about an initial response to infectious agents. The term innate immunity (also called natural or native immunity) refers to the fact that this type of host defense is always present in healthy individuals, prepared to block the entry of microbes and to rapidly eliminate microbes that do succeed in entering host tissues [31].

Adaptive immunity (also called specific or acquired immunity) forms the second line of defence and develops more slowly. It is stimulated by microbes that invade tissues; that is, it adapts to the presence of microbial invaders [31].

2.4 The Human Complement System

The human complement system is a crucial component of the innate immune system that is present over the course of an individual's lifetime. However, it can be recruited and brought into action by the adaptive immune system. It is a system of serum and cell surface proteins that interact with one another and with other molecules of the immune system to generate important effectors of innate and adaptive immune responses. The complement system helps the antibodies (Section 2.5) to clear pathogens from an organism [31, 39].

2.5 What is an Antibody?

An antibody is a large molecule used by the immune system to identify and neutralize foreign objects like bacteria and viruses. The antibody recognizes a unique part of the foreign object, known as the antigen (Section 2.6) [6]. An antibody matches an antigen much as a key matches a lock. Whenever antigen and antibody interlock, the antibody marks the antigen for destruction.

2.6 What is an Antigen?

An antigen is a molecule, usually foreign, that prompts generation of antibodies [31, 48]. An antigen binds to an antibody or a T cell antigen receptor (TCR). Antigens that bind to antibodies include all classes of bio-molecules; for instance, carbohydrates, lipids, and nucleic acids. Most TCRs bind only peptide fragments of proteins complexed with major histocompatibility molecules; both the peptide ligand and the native protein from which it is derived are called T cell antigens.

Although a substance that induces a specific immune response is usually called an antigen, it is more appropriately called an immunogen. Immunogenicity is the ability to induce a humoral and/or cell-mediated immune response. A humoral immune response is one in which host defenses

are brought in by antibodies present in the plasma, lymph, and tissue fluids and it gives protection against extracellular bacteria and foreign macromolecules. A cell-mediated immune response, on the other hand, is one in which the host defenses are brought in by antigen-specific T cells and various other cells which are not specific to a particular target of the immune system and it gives protection against intracellular bacteria, viruses, and cancer [48].

Antigenicity is the ability to combine specifically with secreted antibodies and/or surface receptors on T cells. Although all molecules that have the property of immunogenicity also have the property of antigenicity, the reverse is not true. For instance, some small molecules, called haptens, are antigenic but incapable by themselves of inducing a specific immune response. In other words, they lack immunogenicity [48].

The portion of an antigen that is recognized and bound by an antibody or TCR is known as an epitope [48]. The antigen-antibody (Ag-Ab) interaction is a biomolecular association similar to an enzyme-substrate interaction, except that it does not lead to an irreversible chemical alteration in either the antibody or the antigen. The association between an antibody and an antigen involves various non-covalent interactions such as hydrogen bonds, ionic bonds, hydrophobic interactions, and van der Waals interactions. These interactions are individually weak in comparison to covalent bonds; hence a large number of such interactions are required to form a strong antigen-antibody (Ag-Ab) interaction. In addition, each of these non-covalent interactions operates over a very short distance, usually about 1 angstrom, $\mathring{A}$ (1 x 10^{-7} mm) [48]. As result a strong Ag-Ab interaction depends on a very close fit between the antigen and antibody. Such fits require a high degree of complementarity between antigen and antibody, a requirement that underlies the exquisite specificity that characterizes antigen-antibody interactions [48].

2.7 What is a Vaccine?

A vaccine is a preparation of microbial antigens that is administered to individuals to induce protective immunity against microbial infections. The antigens may be in the form of live but avirulent microorganisms, killed microorganisms, or purified macromolecular components of microorganisms. All these types of vaccines have respective advantages and disadvantages. For instance, live attenuated vaccines induce strong immune response, which often sustain lifelong immunity with just few doses. The disadvantage is that this type of vaccine may mutate to a virulent form. Inactivated or killed vaccines on the other hand are stable and safer than live vaccines; however they result in weaker immune response and they usually require booster shots to provide on-going immunity [31].

Vaccine development has been one of the greatest successes of immunology. Progress in immunology and molecular biology has not only led to the development of effective new vaccines, but it has also led to promising new strategies for finding new vaccine candidates. Knowledge of

the differences in epitopes recognized by T cells and B cells has enabled immunologists to design vaccine candidates to maximize activation of both cellular and humoral immune responses.

2.8 *Neisseria meningitidis*

Neisseria meningitidis causes the meningococcal disease known as cerebrospinal meningitis in developed and developing countries around the world. It is transmitted by person-to-person contact through respiratory droplets of infected people.

According to the World Heath Statistics 2008 [66], the rate of meningococcal meningitis is the highest in the sub-Saharan Africa, which has been named as the "meningitis belt". This belt covers 21 countries with about 350 million population, and the highest disease occurrence is recorded during the dry season. Meningococcal meningitis has hit this area in periodic waves. The last major hit was in 1996/1997 in which more than 220,000 people were affected in 17 countries. After this there was a low disease incidence in the belt until the year 2006, when there was again another rise in the meningitis rates. This increased further in 2007, when there were 54,676 cases of meningitis and 4062 deaths reported in the belt countries [66].

Decrease in immunity among a population against the meningococcus bacterium leads to an increase in the incidence of the disease. In addition, HIV infection, smoking, over-crowding, other respiratory infections, and climatic conditions such as prolonged drought and dust storms contribute to the development of the disease. Meningitis is devastating despite the availability of effective antibiotics. Up to 25% of those who survive are left with permanent damage, such as amputated limbs, mental retardation, hearing loss, speech disorders and paralysis [55].

N. meningitidis is an encapsulated gram-negative bacterium, which infects the upper respiratory tract of ∼10% of the human population and from these infected individuals it occasionally spreads, causing the disease [55]. The onset of symptoms is sudden and death can follow within hours. The flu-like symptoms can rapidly progress into a life-threatening condition, even in healthy individuals [29]. There are two age groups which show higher susceptibility to the disease: children under the age of 2 years and older teenagers. However, during epidemic periods attack rates increase, age distribution broadens (generally 4-19 years of age), and death rates can rise to 10 to 15% [42]. In nonepidemic periods, approximately one of every 100,000 people per year, the bacterium enters the blood stream where it multiplies, causing infections and eventually the disease [55].

Humans are the only hosts which the *Neisseria meningitidis* infects. The bacterium is classified into 13 serogroups: A, B, C, D, 29E, H, I, K, L, X, Y, Z, and W-135 [40]. Among them, five serogroups, A, B, C, Y, and W-135, cause disease [38] and to very minor extent X and Z have been associated with disease [55]. Serogroup A strains cause most of the major epidemics around the world, particularly in less developed countries. Serogroups B and C generally cause endemic disease

and occasionally epidemics. In North America, the most common serogroups causing diseases are C, B, and Y. Serogroup B strains tend to infect children greater than 2 years old, and serogroup C strains are more common in older children, adolescents, and adults. In addition, serogroup C strains are more commonly associated with local outbreaks of the disease. Disease caused by serogroup B and Y strains are usually more sporadic [40].

According to the recent information published by WHO, there are three types of vaccines available for meningococcal meningitis [30]:

- Meningococcal polysaccharide vaccines which are either bivalent (groups A and C), trivalent (groups A, C and W), or tetravalent (groups A, C, Y and W135) have been available to prevent the disease for over 30 years.

- Polysaccharide vaccines for serogroup B cannot be developed due to antigenic mimicry with polysaccharides in human neurologic tissues. As a result, vaccines against B developed in Norway, in Cuba and Netherlands are outer membrane proteins (OMP).

- Since 1999, meningococcal conjugate vaccines (a type of vaccine that is created by joining an antigen to a protein molecule) against group C have been available and widely used. A tetravalent A, C, Y and W135 conjugate vaccine has recently been licensed for use in children and adults in the United States and Canada. In 2001, a partnership was created between WHO and PATH [23] to eradict meningitis in Africa through the development of an affordable meningococcal A conjugate vaccine. From December 2010, the Men A vaccine was introduced nationwide in Burkina Faso, Mali, and Niger. Unlike polysaccharide vaccines, conjugate vaccines are more immunogenic and provide longer immunity.

2.9 fH and fHbp

Activation of the human complement immune response is controlled through membrane-bound and soluble plasma-regulatory proteins, including complement factor H (fH), which is a 155 kDa protein composed of 20 domains (termed complement control protein repeats) [70]. Factor H is the central complement regulator that is found in the plasma and that binds to the surface of host cells and biological surfaces. It acts in combination with additional complement regulators and thus the surface attached factor H assists the membrane anchored regulators. This surface activity is particularly relevant during local inflammation and complement stress, and for damaged host cells [39, 84].

Several pathogens have copied this protective process and adapted the mechanism of sequestering fH to their surface in order to avoid complement-mediated killing. Factor H binding protein (fHbp) is one such antigen which targets factor H, thus enabling the meningococcus to avoid the

innate immune response by inhibiting complement-mediated lysis in human plasma. fHbp is a 27kDa lipoprotein which is present on the surfaces of all strains of *N. meningitidis* and it is the sole receptor for fH on the meningococcus. fHbp is essential for survival of *N. meningitidis* in human blood, in human sera, and in the presence of antimicrobial peptides, hence making it a unique vaccine antigen [68, 70, 84]. Individuals carrying a genetic polymorphism in a presumed regulatory region for factor H have been reported to have increased level of factor H and an increased risk for meningococcal disease [59]. However, one limitation of fHbp as a vaccine antigen is its low density on the surface of some strains of the meningococcus [80] and its sequence variability [42, 55].

Giuliani and et al. [43] have divided the fHbp molecule into a series of regions, termed 'A', 'B', and 'C', in order to study its function. Schneider and coworkers [70] studied these 3 regions and determined that high affinity interactions between fH and fHbp involved all three of these regions; that is, fHbp has an extended recognition site for fH across its entire surface. Schneider and coworkers then identified which of the 20 complement control protein (CCP) domains of fH were involved in the interaction with fHbp. They found that the regions of fH recognized by fHbp were the sixth and seventh domains, CCPs 6 and 7, which they also termed as fH67. They then identified the residues in both fHbp and fH involved (Figure 2.1). Figure 2.1 **a)** illustrates the structure of fHbp that folds to form two β-barrels. The amino-terminal repetitive segment termed as the N-term is part of region A coloured in yellow, region B is coloured in green, and region C in cyan. Figure 2.1 **b)** shows the fHbp-fH67 complex where fH67 is coloured in dark blue. The interaction surfaces between the two molecules forming salt bridges are shown by the red ball-and-stick representations with the inset box as the zoomed illustration of the interactions between fHbp and fH67. Figure 2.1 **c)** shows the topological structural complex of fHbp-fH67. It illustrates the number of the amino acids in fHbp and fH67 that take part in the interactions between the two molecules to form the fHbp-fH67 complex. The fH67 molecule is composed of two parts, CCP 6 and CCP 7 as seen in this part of the figure. The hydrogen bond (H-bond) and the salt-bridge interactions between the residues of the two molecules are shown in red.

Masignani and coworkers [55] classified fHbp into three variants, namely variants 1, 2, and 3 based on sequence diversity. After the classification, it was observed that the amino acid identity between variants 1 and 2 was 74.1%, between variants 1 and 3 was 62.8%, and that between variants 2 and 3 was 84.7% [55].

Using similar analyses as mentioned above, Fletcher and coworkers [42] classified fHbp into two subfamilies, A and B, where subfamily A corresponded to variants 2 and 3 and subfamily B corresponded to variant 1 [42].

When analyzing the fHbp sequence, Beernink et al. [34] divided the sequence into an amino-terminal repetitive element region followed by five variable modular segments. The amino-terminal repetitive region begins with a cysteine (C) residue that was lipidated by signal peptidase II, which

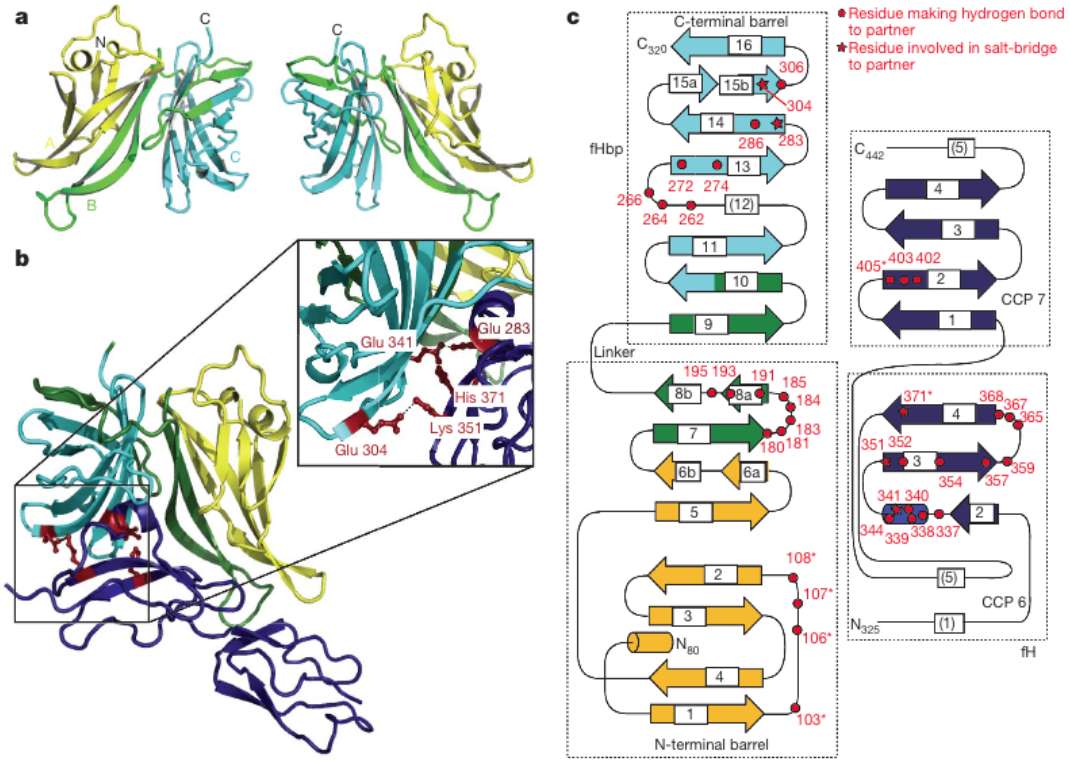


Figure 2.1: Structure of fHbp and its complex with fH67 as shown by Schneider & et al. [70]. **a)** Two views of fHbp (residues 80-320). Region A, yellow; region B, green; region C, cyan. **b)** fHbp-fH67 complex with fHbp coloured as in **a)** and fH67 in dark blue. Side chains from both proteins involved in forming salt bridges across the interaction surface shown in red as ball-and-stick representations. **c)** Topology of fHbp and fH67 coloured as in **b)** with the number of the residues involved in the interactions between the proteins highlighted in red. The numbering scheme used is from UniProtKB sequence Q9JXV4.

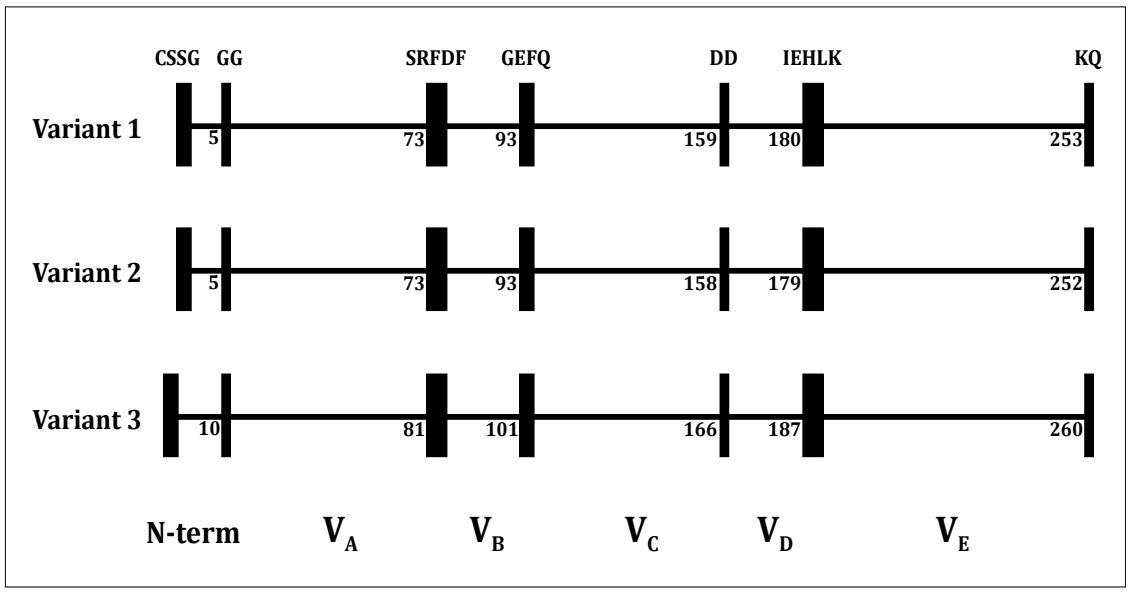


Figure 2.2: Schematic representation of fHbp showing positions of blocks of invariant residues (shown as black vertical rectangles), the amino-terminal repetitive region designated as the N-term, and the variable regions as V_A, V_B, V_C, V_D, and V_E [34]. Masignani and coworkers classified fHbp into three variant groups: Variant 1, Variant 2, and Variant 3 [55].

in turn is followed by three invariant amino acid residues, 'SSG'. This invariant region is followed by a repetitive element consisting of 1 to 6 glycines (G) and / or serines (S) followed by two invariant glycine residues (Figure 2.2).

The five modular variable regions have been termed Regions V_A, V_B, V_C, V_D, and V_E as shown in Figure 2.2. These five regions are separated by blocks of invariant residues. For instance, Region V_A starts right after the invariant region 'GG' of the amino-terminal region and extends until the invariant segment, 'SRFDF' after which is the region V_B. Hence the variable segments are each separated by the blocks of invariant segments (Figure 2.2). These five modular variable segments have specific numbers of amino acids as mentioned by Beernink et al. [34] (Figure 2.2).

- V_A contains 66 to 69 amino acids.
- V_B contains 15 amino acids.
- V_C contains 61 to 62 amino acids.
- V_D contains 19 amino acids.
- V_E contains 68 to 71 amino acids.

2.10 Evolution & Mutation

Evolution is the process by which living things change through a process of incremental transformation over extended periods of time. It is a modification which occurs due to genetic variability, which in turn is produced by the process of mutation. In the early days of genetic studies, any genetic change in phenotypic characteristics was referred to as mutation, with no knowledge of the reason behind the change. However, with time and with further studies and research it was revealed that various factors are involved in causing genetic changes of phenotypes. These factors can be studied at three different levels, namely, molecular, chromosomal, and genomic [60].

For this research, it is only required to know what evolution is at the molecular level. A simple definition of molecular evolution is that it is the process of evolution at the scale of DNA, RNA, and proteins, and that mutation plays an important role in it [8]. Mutation is a permanent, transmissible change to the genetic material (usually DNA or RNA) of a cell and it is considered a driving force of evolution.

DNA and RNA with viruses are the carriers of all genetic information that controls the morphological and physiological characteristics of organisms. Any mutational changes in these characteristics are hence the result of some changes in DNA molecules and RNA with viruses. There are four basic types of changes in DNA and RNA with viruses: (i) substitution of a nucleotide for another nucleotide (Figure 2.3 b), (ii) deletion of nucleotides (Figure 2.3 c), (iii) insertion of nucleotides (Figure 2.3 d), and (iv) inversion of nucleotides, where nucleotides are reversed in a sequence (Figure 2.3 e). Insertion, deletion, and inversion may occur with one or more nucleotides as a unit. Nucleotide substitutions are of two types: (i) transition, which is the substitution of a purine (adenine or guanine) for another purine or the substitution of a pyrimidine (thymine or cytosine) for another pyrimidine, and (ii) transversion, which is the substitution of a purine for pyrimidine or vice versa [60].

The amino acid sequence of a protein is determined by the nucleotide sequence of its gene. Hence any change in amino acid sequence is caused by a mutation occurring in DNA. However, a mutational change in DNA does not always result in a change in amino acid sequence because of the degeneracy of the genetic code (synonyms of codons). Mutations that result in synonymous codons are called synonymous or silent mutations, while others are called nonsynonymous or amino acid altering mutations [60].

The other important aspect of molecular evolution is determining regions which are under positive selection pressure and those that are under negative selection pressure. A positive selection site is a site which is under active natural selection; that is, nature preferably changes or mutates that position in order to bring new evolutionary results [56]. For instance, over time this process can result in the emergence of new mutants which have higher fitnesses than the average in the

population (provided the environment is also optimal for the change), and the frequencies of the mutants increase in the following generations [75]. In other words, positive selection is an important process by which evolution takes place within a population of organisms. A negative selection site, on the other hand, is the opposite of positive selection, where nature conserves the type of amino acid present at that site and does not want changes to occur. In fact it will prevent any changes from occurring to that site [56]. As a result of this process, newly produced mutants have lower fitnesses than the average in the population, provided the environment is also changing. In that case, conservation will likely be deleterious. However, if the environment has not changed, and if the sequence is already somehow "optimal", then conserving the sequence is beneficial to the survival of the organism. In addition to all these, negative selection pressure leads to a decrease in the frequencies of the mutants in the following generations [75].

In a protein sequence set, positive and negative selection sites can be determined by finding ratio of the number of nonsynonymous substitutions per nonsynonymous site (designated as K_A or d_n) to the number of synonymous substitutions per synonymous site (designated as K_S or d_s) [56]. The rate of synonymous substitution is much higher than that of nonsynonymous substitution and is similar for many different genes of closely related species [61]. Methods for estimating the K_A/K_S ratio can be classified into two groups: approximate method for closely related sequences and maximum-likelihood method for distantly related sequences [82].

The approximate method involves three steps [82]:

1. counting the number of synonymous and nonsynonymous sites in the two sequences—usually by multiplying the sequence length by the proportion of each class of substitution
2. counting the number of synonymous and nonsynonymous substitutions
3. correcting for multiple substitutions

The maximum-likelihood approach uses probability theory to complete the above three steps simultaneously [82].

Although the basic genetic change occurs in DNA, evolutionary change of proteins is also important, since proteins are essential building blocks for morphological characteristics and carrying out physiological functions. Also they are more conserved than DNA sequences and thus provide useful information on long-term evolution of genes or species [62].

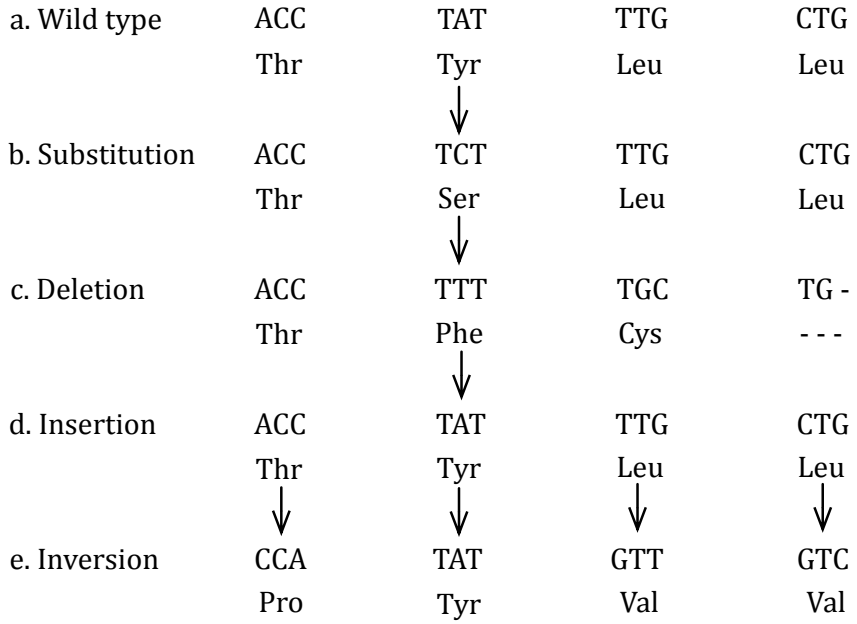


Figure 2.3: Four basic types of mutations at the nucleotide level. The nucleotide sequence is represented in units of codons or nucleotide triplets in order to show how the encoded amino acids are changed by the nucleotide changes [60].

2.11 Mature Protein Part and Signal Protein Part

The "mature protein" part, also termed as the "final peptide", is the protein product following post-translational modification of a protein [18]. Post-translational modification is the chemical modification of a protein after its translation step, which is one of the later steps in protein biosynthesis [9].

A signal peptide is a short (3-60 amino acids long) peptide chain that directs the transport of a protein. "Targeting signal", "signal sequence", "transit peptide", or "localization signal" [11] are some of the synonyms for the term "signal peptide".

2.12 Protein Structure Prediction

One of the major goals of Bioinformatics is to understand the relationship between amino acid sequence and three-dimensional structure in proteins. If this relationship were known, then the structure of a protein could be reliably predicted from the amino acid sequence. Knowing the structure of a protein is important, since it provides much more biological information than the protein amino acid sequence. The structural analysis of proteins plays a significant role in detecting many physical-chemical properties of proteins and prediction of protein-protein interactions [57].

The structure of a protein is conventionally described in four ways. The *primary* structure of a protein is the sequence of amino acids produced at ribosomes. The *secondary* structure of a protein describes those parts of the primary structure that fold into regular and repeated patterns, such as α-helices, β-sheets, or turns. The *tertiary* structure results from interactions between secondary structure elements, forming more complex units and providing the three-dimensional shape of the protein. The *quaternary* structure of a protein is a description of how several separate polypeptide sequences have come together to form a complex protein [46].

2.13 Energy of Protein Structures

Energy characteristics are distributions of potential energy over amino acid chains of proteins. Conformational changes, mutations, structural deformations and other disorders are reflected in energy distributions. Hence energy characteristics can be used in detection and verification of such states [58].

Minimization of energy is an important concept used in repairing distorted geometrics of a structure of protein, especially if the protein has some kind of mutation, when it is distorted manually, or when there is a manual reconstruction of loops in the structure. Energy minimization can repair distorted geometrics by moving atoms to release internal constraints [20].

Energy plays a significant role in evaluating a structure of a peptide molecule. For instance, when predicting a tertiary structure of a protein using a template structure that is already available, the energy that is determined during this prediction plays an important role in determining how accurately the structure has been predicted. It is usually seen that predicted structures with higher percentage of structural identity with the template structure will exhibit lower energy values, whereas the opposite is true when structures have greater structural variations from the template structure. Energy of peptide molecules are completely based on their structural properties and not on their sequential properties. Two peptide molecules sequentially similar can have vast structural differences due to conformational changes in their loops or side chains, etc. thus exhibiting pronounced changes in energy values.

Chapter 3

Research Goal

Evolution does not happen at random and one strain of a bacteria cannot evolve directly to another arbitrary strain. Nature allows only certain paths in going from one strain to another and such transformation is influenced by both past mutation and current genetic diversity. Based on this idea, this research aims to identify what the fHbp antigen of *Neisseria meningitidis* bacterium is likely to mutate into so that vaccines and therapies can be developed in advance.

The factor H binding protein (fHbp) was chosen for this research because it is present on the surface of all strains of *N. meningitidis* and plays a significant role in the survival of the bacterium in human blood in the presence of antimicrobial peptides, hence making it a potential vaccine antigen. The goal of this research is to predict valid new variants of fHbp so that vaccines and therapies can be developed in advance before the mutants appear and cause the disease. The variants generated are to have the same sequential and structural properties as those of the original fHbp sequences but not be duplicates of the existing ones.

Creating new variants of fHbp and validating whether they have the same properties as those of the existing ones is not a single-step process. The followings are the specific objectives of this research in order to meet the overall goal.

- Construction of a methodology which is extensible in nature, such that it will form the foundation for additional work.
- Determination of the characteristics of the existing fHbp sequences.
- Characterization of the constraints between the amino acid positions with variations in the fHbp sequence alignment.
- Selection of a suitable protein modelling software for predicting the tertiary structures of the sequences.
- Selection of a suitable software for determining the energy of the tertiary structures.
- Development of a pipeline of programs for generating the new variants of fHbp. The program should be both extensible and flexible in nature and should be based on the properties of the existing fHbp sequences. It can be used as a prototype for additional future work.

- Determination of the validity of new variants based on the properties of the existing fHbp sequences.

When the overall goal of the research is met, the viability of the variants can be determined in a wet lab; that is, whether the *N. meningitidis* bacterium with the proposed altered fHbp antigen can exist in nature or not. This can be carried out as a future work for this research.

Chapter 4

Data & Methodology

This section discusses the data collected and the methodology followed for this research. However, parts of the results of Chapter 5 are also discussed in this section since some of the steps in the methodology were dependent on the results generated in the previous steps. This has been done to enhance reader comprehension of each step of the methodology.

4.1 Data

The data—that is, the factor H binding protein sequences for this research—were downloaded from the UniProtKB database [27]. The sequences were retrieved using 'fHbp', 'gna1870', or 'nmb1870' as the query keyword since all of these are synonyms for factor H binding protein. The query was further constrained by specifying that the source organism has taxonomy identifier 487, that for *Neisseria meningitidis*. In total 270 matching sequences were downloaded on June 09, 2009. There were 197 sequences of fHbp, 72 of gna1870, and 1 sequence of nmb1870.

4.2 Methodology

The methodology followed for this research is illustrated in Figure 4.1 with explanations for each step in the sections that follow.

Step 1

For this research, all known fHbp sequences independent of serotype and serogroup were downloaded from UniProtKB. Other aspects of this step have already been described in Section 4.1.

Step 2

There are duplicates and fragments present in protein sequence databases [32]. One of the prime reasons for this may be that removing redundancy based on genome annotation (Section 2.1) is a difficult task. However, removing duplication based on sequence identity is a tractable problem. Publicly accessible software exists for doing it with high degree of correctness. There is far less software for removing redundant sequences based on annotation and it gives only mixed results.

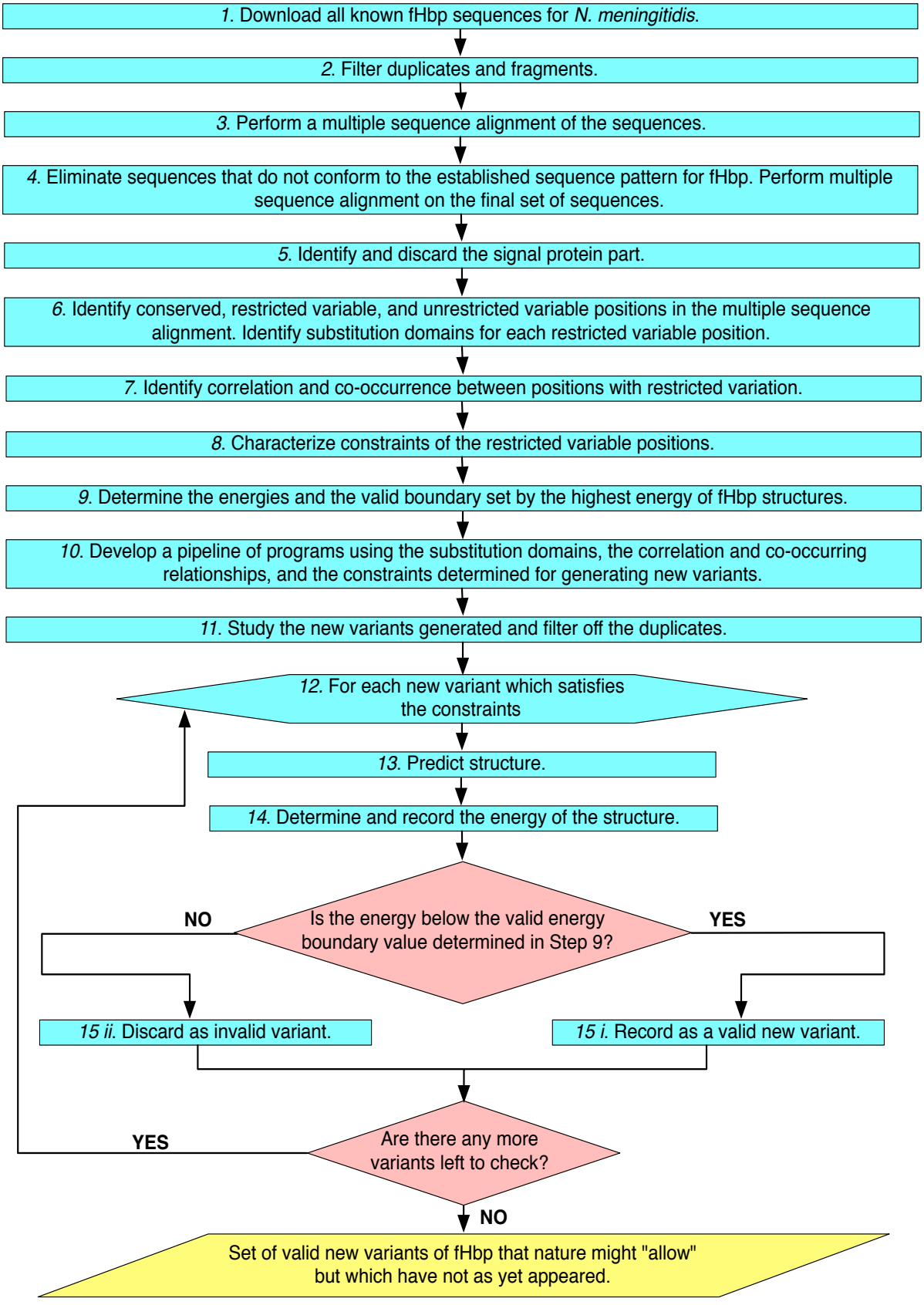


Figure 4.1: Methodology Flow Chart

At the databases which filter for redundancy, only one type of redundancy filtering is performed. For instance, UniProtKB filters only by annotation; UniRef filters only by sequence identity. No database combines these two approaches, for example filtering by both annotation and sequence identity. Hence, for this research the strategy was to download from UniProtKB to take advantage of the filtering based on annotation which they provide and then filtering by sequence identity in-house.

For this research, duplicates and fragments were filtered using an extremely fast protein sequence clustering program called CD-HIT (Cluster Database at High Identity with Tolerance) [14, 52, 53, 54]. CD-HIT uses a greedy incremental clustering algorithm that is capable of handling large databases like NCBI's non-redundant protein database containing more than 560 000 sequences. The algorithm was developed by Weizhong Li in 2001 and is still under active development, with promises of new features in future. CD-HIT is used at many research and educational institutions. For instance, it is used at UniProt (Universal Protein Resource) to generate the UniRef (UniProt Reference Clusters) clusters at 100%, 90%, and 50% sequence identity thresholds and also at PDB to remove redundant sequences [52, 53, 54]. CD-HIT has been hosted at bioinformatics.org as an open source project since 2004 [14].

Using CD-HIT to remove duplicates and fragments.

The three sets of sequences downloaded from UniProtKB were merged into one set, totaling 270 sequences. CD-HIT was then used at 100% sequence identity on this merged set and any duplicates or fragments identified were removed. This resulted in a set of 200 unique fHbp sequences.

Step 3

This step involved the multiple sequence alignment (Section 2.2) of the 200 sequences obtained from the previous step using ClustalW. ClustalW was chosen because it is an easy-to-use general purpose multiple sequence alignment program for DNA or proteins which implements a straightforward progressive alignment algorithm. A progressive alignment algorithm is one in which sequences are added one by one in order, starting with the most similar sequences. It takes a modest amount of time and memory to align a set of 200 sequences of typical length. ClustalW produces biologically meaningful multiple sequence alignments of non-divergent sequences, as in this case [19, 65].

After the multiple sequence alignment was generated by ClustalW, the EMBOSS *prettyplot* program was used to represent it [5]. This was because *prettyplot* displays aligned sequences with colouring and boxing, making it easier to locate conserved and variable regions. It reads in a set of aligned DNA or protein sequences (e.g. the output from ClustalW) and displays them graphically, with conserved regions highlighted in various ways.

Step 4

Close study of a sequence alignment facilitates detection of any sequence that may seem erroneous or inappropriate. For example, sometimes sequences belonging to a different species may be inadvertently retrieved from a sequence database. Erroneous or inappropriate sequences should be discarded at an early stage of the methodology so that they do not adversely affect the results.

Table 4.1: Result generated after investigation of the 200 sequences.

Length of the sequences	**Number of sequences**
84 amino acids	1 sequence
133 amino acids	1 sequence
250-260 amino acids	125 sequences
261-270 amino acids	19 sequences
271-280 amino acids	38 sequences
281-290 amino acids	12 sequences
320 amino acids	1 sequence
427 amino acids	1 sequence
492 amino acids	1 sequence
497 amino acids	1 sequence

In this step of the research, the multiple sequence alignment of the 200 sequences was studied to identify any sequences that did not conform to the signature criteria of fHbp: having the amino-terminal repetitive element region and five modular variable segments flanked by invariant regions [34] as discussed in Section 2.9. The first task was to find the lengths of the 200 sequences in the merged set. Table 4.1 lists the results. All the sequences were then closely studied to determine if any were problematic and should be discarded.

Investigating the sequence of length 84 amino acids.

Referring to Table 4.1, there was one sequence with a length of 84 amino acids. Investigation of the sequence determined that it violated the sequential structure of fHbp sequences as specified in Section 2.9. Mention of the term 'fHbp' in the reference portion of its annotation had resulted in it being retrieved. Hence it was discarded from the set of fHbp sequences.

Investigating the sequence of length 133 amino acids.

The sequence with 133 amino acids in Table 4.1 was too short according to the allowed sizes of variable segments mentioned in Section 2.9. Further study showed that this sequence had the amino-terminal repetitive region, regions V_A and V_B, but had an incomplete region V_C and none

of the regions V_D or V_E, confirming that it was an incomplete sequence. Hence this sequence was eliminated from the final set of fHbp sequences.

Investigating the sequence of length 320 amino acids.

In order to investigate whether the sequence with 320 amino acids could be kept in the final list of sequences, the results generated by CD-HIT at different sequence identities were examined. It was found that at 95% sequence identity, 24 sequences clustered with this particular sequence. Thus the sequence had very high similarity with other sequences in the set. However, there was no clustering at 100% sequence identity, since all the duplicates were removed in Step 2. In addition, this sequence had the amino-terminal repetitive region plus all five of the variable segments including the invariant regions separating the variable segments. Also the size of the variable segments were as specified in Section 2.9. Hence it was kept in the final list of fHbp sequences.

Investigating the three sequences of length 427, 492, and 497 amino acids.

Of the three longest sequences in Table 4.1, none formed clusters with any other sequences at 100, 95, 90, 85, or 60% sequence identity when CD-HIT was used. However, the sequences clustered with each other. When these sequences were investigated individually, it was found that they had a missing invariant 'SSG' region after cysteine (C) in the amino-terminal repetitive segment. Instead they had 'GGG'. Further studies showed that they had none of the invariant segments that delimit the five variable regions (Figure 2.2). Upon checking these sequences at the UniProtKB website it was found that the proteins had 'fHbp' in their annotation: *"fHbp, nadA and gna2132 in group B meningococci"*. This explained their presence in the retrieved set of sequences. Based on these findings the three sequences were eliminated from the set of sequences.

Removal of more sequences.

The sequences were further probed to determine if there were any which did not fit the criteria of having the amino-terminal repetitive region, the invariant regions flanking the variable segments with the specified number of amino acids in the variable regions as mentioned in Section 2.9. Four sequences were found to be variant in the amino-terminal repetitive region (Figure 2.2). Normally, these sequences should have an invariant 'CSSG' followed by one to six glycines or serines, followed by the invariant 'GG'. Instead, one sequence had 'GGGSGGGG' (i.e. 8 glycines or serines) between the two invariant regions. Another had a sequence with 4 extra glycines or serines (10 in all) between the invariant regions. The third sequence had an 'R' instead of a 'G' in the first invariant; that is, it had 'CSSR' instead of 'CSSG'. The fourth sequence had 'CSG' instead of 'CSSG' as the first invariant region. Based on these observations the four sequences were discarded from the set of sequences.

One of the sequences had a missing 'Q' in the invariant segment 'KQ' at the end of the region V_E; it only had 'K' instead of 'KQ'. Hence, this sequence was eliminated.

In total this step discarded ten sequences from the set of sequences. As a result the final set had 190 fHbp sequences. Multiple sequence alignment using ClustalW was performed again on this final set and the EMBOSS *prettyplot* program was used to represent it (as was done in Step 3).

Step 5

In this step, the consensus sequence for the multiple sequence alignment was determined. Then, for each sequence, the mature and signal portions (Section 2.11) were identified, and the signal portion discarded.

A consensus sequence is a single sequence that represents the most common nucleotide or amino acid found within corresponding columns of a multiple sequence alignment [57]. The consensus sequence for this research was created from the multiple sequence alignment using the *cons* program from EMBOSS [4].

For this research, SignalP 3.0 Server [17, 35, 63] was used for determining the mature part of the sequences. SignalP 3.0 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms, including gram-negative prokaryotes. The method uses a prediction model based on a combination of several artificial neural networks and hidden Markov models [17, 41]. SignalP 3.0 Server suggests the usage of the consensus sequence as input for predicting the signal protein part and the mature protein part.

From the published work of Masignani et al. [55], it was already known that the mature part of the fHbp protein sequence starts from the invariant 'CSSG' region. For further confirmation, SignalP 3.0 Server [17] was used in this research to determine the mature protein part. From Section 2.6, it was already known that Beernink et al. [34] had divided the fHbp sequence into various segments where the first part is the amino terminal repetitive segment, termed as the "N-term" in Figure 2.2. Hence, the consensus sequence up to and including the amino terminal repetitive region, that is, up to the "N-term" was input to the SignalP 3.0 Server [17]. In response, the server reported that the mature part of the consensus sequence of the 190 fHbp sequences started at the cysteine of the amino-terminal repetitive region; that is, it started from the invariant region 'CSSG'. This was consistent with the finding of Masignani et al. [55]. For further confirmation, a couple of the 190 fHbp sequences themselves were input to the SignalP 3.0 Server. The results generated showed the same output like that of the consensus sequence of the 190 fHbp sequences mentioned above. Hence, the portion upstream of the invariant 'CSSG' region was then discarded from each sequence in the collection. This resulted in a set of 190 sequences with their mature protein part only.

Step 6

In a set of homologous protein sequences, there are regions which are conserved and those which are not conserved (variable). The latter regions can be further divided into two groups: segments that are restricted, meaning that there is little variation or the variation is very limited, and segments that are unrestricted, meaning there is a high degree of variation in these regions. Identification of conserved and variable regions is needed to determine the correlation or co-occurrence relationships between positions of a homologous set of sequences. Thus, at this step the conserved, restricted variable, and unrestricted variable regions were identified in the multiple sequence alignment.

Five scripts (given in Appendix A) were developed to automate this step. These scripts (Appendix A.4 and A.5) determined which columns were conserved and which were not, and also whether variation was restricted or unrestricted. They also determined which columns had 'gap' symbols (Appendix A.3) and the domain of possible amnio acid substitutions for each restricted variable position (Appendix A.1 and A.2).

Step 7

For the purpose of this thesis, "correlation" is taken to have a particular meaning. Here, correlation between different positions in a sequence alignment indicates to what extent a mutation in one position of an alignment affects the presence or absence of other amino acids in another position; that is, correlation for this thesis is directional. A co-occurring relationship, on the other hand, means that the peptides in different positions occur together in a pattern. Two types of co-occurring relationships were observed in this research: "fully" co-occurring and "partially" co-occurring. Figure 4.2 illustrates the concepts of correlation and both types of co-occurring relationships. In the figure, amino acid 'D' in position 3 is correlated to the amino acid 'N' in position 1, both of which are highlighted in red for convenience. Whenever 'D' occurs in position 3 there has to be an 'N' in position 1. However, the reverse is not true. When there is an 'N' (highlighted in light green) in position 1 there may be other amino acids in position 3, for instance, 'E' and 'Q'. Hence, amino acid 'D' in position 3 is correlated to amino acid 'N' in position 1. Positions 2, 4, 5, and 6 depict a co-occurring relationship among their amino acids, with position 5 illustrating a "partially" co-occurring relationship and positions 2, 4, and 6 a "fully" co-occurring relationship. It is seen that amino acids 'A', 'V', 'L' (all highlighted in yellow) and amino acids 'S', 'R', 'F' (all highlighted in green) in positions 2, 4, and 6, respectively, occur in coordinated, fixed manner. Thus positions 2, 4, and 6 are "fully" co-occurring. Amino acids 'C/I' (highlighted in yellow) and 'M' (highlighted in green) in position 5 occur in coordination with amino acids of positions 2, 4, and 6, with amino acids 'C/I' constituting the same proportion as the amino acids 'A', 'V', and 'L' in positions 2, 4, and 6. Since there are more than one type of amino acids ('C/I')

Position 1	Position 2	Position 3	Position 4	Position 5	Position 6
G	A	E	V	C	L
G	A	E	V	I	L
N	A	E	V	I	L
N	A	D	V	C	L
N	A	Q	V	C	L
G	S	Q	R	M	F
G	S	Q	R	M	F
N	S	D	R	M	F
G	S	E	R	M	F
G	S	E	R	M	F
N	A	D	V	C	L
N	A	Q	V	I	L

Figure 4.2: Correlation and Co-occurring relationships

constituting the same proportion as the amino acids 'A', 'V', and 'L', the position with more than one type of amino acids is "partially" co-occurring with the positions with single type of amino acid. Thus position 5 is "partially" co-occurring with positions 2, 4, and 6. In general, it can be said that "co-occurrence" is a more restrictive relationship than "correlation".

For this research, different positions of the aligned sequences were visually examined to determine if variations in one position of one of the five restricted variable regions were correlated or co-occurring with changes in another position (which may be part of the same region or a different one).

Step 8

This step characterizes constraints of the restricted variable positions of the existing fHbp sequence alignment. The constraints can be molecular evolutionary constraints, structural constraints, etc. Constraints help to determine conditions, such as whether one amino acid will have higher possibility of substituting another amino acid or if mutation between two amino acids involves presence of an intermediate amino acid, etc. For this research, determining the constraints are important, since they act as templates for generating valid new variants in the later steps.

The following are some constraints, some of which were utilized in this research. The remaining constraints couldn't be utilized since they fell outside of the scope of this thesis. However, they can be carried on as part of future work (Section 6.2).

i. Molecular Evolution. This constraint is based on the evolution or change from one amino acid to another. The mutation may involve one codon position or more than one. Some

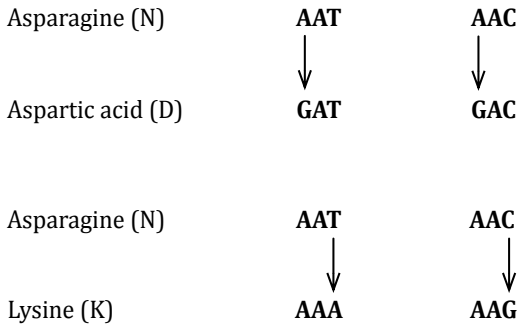


Figure 4.3: Mutating from asparagine (N) to aspartic acid (D) and from asparagine (N) to lysine (K) requires nucleotide changes in only one codon position.

mutations require a third–temporary and intermediate–amino acid to which mutation may occur before going to the final changed amino acid. Amino acid mutations which require changes in only one codon position are more likely to occur than those which require more than one or involve an intermediate residue. Another relevant characteristic of molecular evolution is the distinction between transition and transversion (Section 2.7). Transition is more likely to occur than transversion because the former involves mutational changes between the same types of bases; for instance, between the two purine bases or between the two pyrimidine bases. The later involves mutational changes between the two different types of bases, purines and pyrimidines.

The following examples illustrate the ideas above.

Example 1. Mutating from asparagine (N) to aspartic acid (D) and from asparagine (N) to lysine (K) requires nucleotide changes in only one codon position as illustrated by Figure 4.3.

For both the evolutions, change is required in only one codon position. Despite this, according to the properties of the amino acids illustrated in *"Amino Acid Properties and Consequences of Substitutions"* in the book *"Bioinformatics for Geneticists"* [36], the likelihood of mutating from asparagine (N) to aspartic acid (D) is greater than that from asparagine (N) to lysine (K); that is, substitution of N by D is more favored than substitution of N by K. In addition, the evolution from N to D is a transition since the nucleotide change has been between two purines (adenine and guanine) whereas the one between N to K is transversion involving change from pyrimidine to purine (thymine to adenine and cytosine to guanine). Hence, this example illustrates the constraint where one amino acid has higher possibility of substituting for another amino acid.

Example 2. Suppose asparagine (N) is substituted for glutamine (Q). There are two codons for N and two for Q. It is possible for both the codons for N to mutate to a codon for Q by

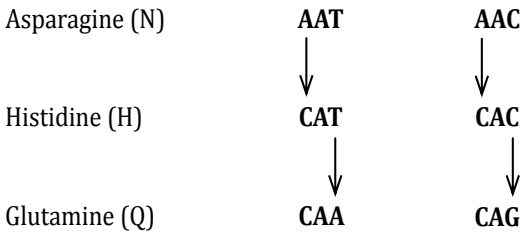


Figure 4.4: Mutating from asparagine (N) to glutamine (Q) involves mutating to an intermediate codon for histidine (H).

involving an intermediate codon for histidine (H), with one nucleotide substitution resulting in the codon for H and a second yielding the codon for Q. This is illustrated in Figure 4.4.

Since the mutation from N to H involves change in only one nucleotide and that to Q involves two nucleotides, there is a high possibility of N mutating to H before it mutates to Q; that is, mutating to an intermediate amino acid before mutating to the final amino acid. In addition, according to the substitution properties illustrated in *"Amino Acid Properties and Consequences of Substitutions"* [36], substitution of N by H is more favoured than substitution of N by Q. Hence, this example illustrates the constraint where the mutation from one amino acid to another may involve an intermediate amino acid.

The positive and negative selection sites in a sequence alignment is another aspect leading to significant constraints on molecular evolution (Section 2.7 discusses positive and negative selection pressure); for instance, positive selection sites would be more likely to change, and amino acids in these positions would be more prone to mutation, thus changing to different amino acids. The negative selection sites, on the other hand, would tend to be more conserved. Also, these sites would more likely be binding sites of a protein molecule, since the binding sites tend to remain unchanged, so that they would be recognized by the molecule with which they bind.

Codon-based detection of positive and negative selection involves a method known as the sliding window approach. It requires two parameters: a window width and a step length. The window width is the size of the range of positions in a sequence alignment that is taken into account for determining the K_A/K_S ratio for those positions. Step length is the number of positions that the window is moved along the alignment in order to determine the next K_A/K_S ratio [81]. If the ratio of K_A/K_S is greater than 1, then the positions within that window width are under positive selection pressure and if it is less than 1, then they are under negative selection pressure [56]. There are many tools which determine the ratio of K_A/K_S; for instance, MEGA4 [22, 76], WSPMaker [51], and KaKs Calculator 2.0 [2, 83].

ii. Spatial and physicochemical constraints. This constraint is based on the tertiary structures of the sequences and the physicochemical properties of the amino acids. The tertiary structural constraints would involve two things:

(a) amino acids that must be within certain distances of each other to form, for example, an active site;

(b) amino acids with conflicting physicochemical properties that cannot be within certain proximities of each other.

The tertiary structures can be predicted using various protein modelling software which can predict the structures and can also provide information about protein folding, interactions, etc. Some widely used protein modelling servers involved in tertiary structure prediction are SWISS-MODEL [13, 33, 47, 67], CPHmodels [15, 64], and EsyPred3D [21, 50], all of which are good at predicting structures. Also, the protein structures can be superimposed on each other in order to determine the differences in their structures and to calculate their respective root mean square deviation (termed RMSD). RMSD is the measure of the average distance between the backbones of superimposed protein structures and is usually measured in Angstroms, $\mathring{A}$ [10]. The higher the RMSD value, the more deviated the structures are. There are various molecular modelling and visualization tools which can be used to view the tertiary structures; for instance, Swiss-PdbViewer [20, 45], Rasmol [26, 69], Jmol [12], and PyMol [25, 71].

iii. Sampling frequency. This constraint is based on the amino acid combinations that are not observed in a sequence alignment. This is an important but the most difficult category to deal with because it means that a combination of amino acids is not taken into consideration since it is not seen in the collection of the sequences that has been gathered. Not seeing a specific correlation or co-occurrence of certain amino acids does not mean it does not occur at all in nature. There can be instances where certain positions must have specific amino acids together but the combination was not considered because it was not seen in the sequences collected for a research. Hence, the constraint here involves taking into consideration all the allowed combinations of amino acids either correlated or co-occurring with each other.

For this research, the purpose of this step was to determine the constraints on the restricted variable positions. Hence, the three constraints characterized above can be used while generating the new variants of fHbp. Below described are the characterized constraints and how some of them have been implemented in this research.

i. Molecular Evolution. Finding amino acids where one has higher possibility of substituting another amino acid and finding intermediate residues in possible mutations was deemed out

of the scope of the thesis, hence this aspect of molecular evolution was not included in this research work. However, it can be considered as a prospective future work for this research as discussed in Section 6.2 (Future Work).

Determining the ratio of K_A/K_S for detecting positive and negative selection sites was included in this research. This was done in order to establish the templates for generating valid new variants. The determination of K_A/K_S ratio required nucleotide sequences of the protein sequences. Hence, each corresponding nucleotide sequence of the 190 fHbp sequences were downloaded from UniProtKB. Since our 190 fHbp sequences were closely related, the approximate method was used for determining the K_A/K_S ratio. The tool used for this purpose was MEGA4 [22, 76]. MEGA4 has in total five methods by which the K_A/K_S ratio can be calculated. The most recent method is the Kumar Method, which is a modification of previous methods [62]. In addition, this method takes into account transitions and transversions between codons while calculating K_A/K_S. For these reasons, the Kumar Method was chosen. The sliding window approach was used, where the window width was 18 base pairs and the step length was 3 base pairs. K_S was estimated from the whole sequence alignment of the 190 sequences. This was because K_S tends to 0 for closely related sequences (like the 190 fHbp sequences of this research) and/or when the window size is small [81].

ii. Spatial and physicochemical constraints. SWISS-MODEL [13, 33, 47, 67] was used for predicting the tertiary structures of the 190 fHbp sequences. SWISS-MODEL, a fully automated protein structure homology-modeling server, is a web-based integrated service that assists and guides in building protein homology models at different levels of complexity. For visualization, analysis, superimposition between two structures and for RMSD calculations Swiss-PdbViewer (also known as DeepView) [20, 45] was used.

The mutationally correlated or co-occurring positions in region V_A were studied in the tertiary structures generated by SWISS-MODEL and Swiss-PdbViewer. Positions that were spatially close to this region were taken into consideration. This resulted in examining not only positions within region V_A, but also segments from the other regions (V_B - V_E). It was observed that even though residues were far apart in the sequential alignment, they could be in close spatial proximity in the tertiary structures and have a correlational or co-occurring relationships between them. It was out of the scope of this thesis to consider the second tertiary structural constraint mentioned above: "amino acids with conflicting physicochemical properties that cannot be within certain proximities of each other". However, in future, this constraint can be considered.

iii. Sampling frequency. This research was strictly based on the sequences that were collected from UniProtKB (Section 4.1) and the combinations of amino acids considered were from the

alignment of those sequences. Hence, the constraint involving taking into consideration all the possible combinations of amino acids either correlated or co-occurring with each other was not utilized for this thesis. Correlations, co-occurrences and the domain of possible amino acid substitutions for each position were taken into considerations only if they were seen to occur in the sequences downloaded for this research.

Step 9

Energy of protein structures reflects conformational changes, mutations, and structural deformations and other disorders [58]. Section 2.9 briefly outlines the importance of the energy of protein structures. There are various tools to determine the energy of protein structures and FoldX is a widely-used one. It is a computer algorithm developed at the European Molecular Biology Laboratory in Heidelberg. FoldX uses a full atomic description of the structures of proteins [44, 72, 73, 74]. The FoldX energy function is as follows [44, 72, 73, 74]:

$\Delta G = W_{vdw} \bullet \Delta G_{vdw} + W_{solvH} \bullet \Delta G_{solvH} + W_{solvP} \bullet \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{Kon} + W_{mc} \bullet T \bullet \Delta S_{mc} + W_{sc} \bullet T \bullet \Delta S_{sc}$

- ΔG : free energy of unfolding of a target protein
- ΔG_{vdw} : sum of the van der Waals contributions of all atoms with respect to the same interactions with the solvent
- ΔG_{solvH} : difference in solvation energy for apolar group when they change from the unfolded to the folded state
- ΔG_{solvP} : difference in solvation energy for polar group when they change from the unfolded to the folded state
- ΔG_{wb} : extra stabilising free energy provided by a water molecule making more than one hydrogen bond to the protein (water bridges) that cannot be taken into account with non-explicit solvent approximations
- ΔG_{hbond} : free energy difference between the formation of an intra-molecular hydrogen bond compared to inter-molecular hydrogen-bond formation (with solvent)
- ΔG_{el} : the electrostatic contribution of charged groups, including the helix dipole
- ΔS_{mc} : entropy cost of fixing the backbone in the folded state and this is dependent on the intrinsic tendency of a particular amino acid to adopt certain dihedral angles
- ΔS_{sc} : entropic cost of fixing a side chain in a particular conformation
- ΔG_{Kon} : reflects the effect of electrostatic interactions on the association constant K_{on} (this applies only to the subunit binding energies)

Figure 4.5: Modular structure of the pipeline of programs

- W_{vdw}, W_{solvH}, W_{solvP}, W_{mc}, and W_{sc} (van der Waals contribution), which is 0.33 (the van der Waals contributions are derived from vapor to water energy transfer, while in the protein it derived when going from solvent to protein).

From the previous step we had the tertiary structures of all the 190 fHbp sequences. Now as the first stage of this step, energy minimization was performed on these 190 tertiary structures. The energy minimization was performed using Swiss-PdbViewer [20, 45], which repairs distorted geometrics by moving atoms [20, 45]. Next, FoldX was used on the structures with minimized energy and energy values were calculated. From these energy values the highest energy value was identified in order to set the boundary for the valid energy values that fHbp sequences can have.

Step 10

In this step a pipeline of programs was developed using the domains of possible amino acid substitutions for each position (Step 6), the correlation and co-occurrence relationships (Step 7) and all the constraints determined (Step 8) for generating new variants. The programs were written in the Perl programming language and were coordinated using a UNIX shell script written in bash. Each program had different functionality in setting the amino acids for the different positions in the alignment. There were programs which only generated amino acids for the positive selection sites, programs which generated amino acids for the amino-terminal repetitive regions, programs for the conserved regions and for the restricted variable regions, etc. Algorithm 4.1 to Algorithm 4.6 illustrate some of the programs implementing these various functions. A design goal of the pipeline was flexibility in the order of execution of individual programs; that is, the order of the Perl scripts within the pipeline could be varied. The exception was the first program in the pipeline. To achieve this design goal each program was made independent and able to use outputs generated from any program preceding it in the pipeline. Another characteristic was that additional Perl scripts (e.g. implementing additional constraints) can be added to the pipeline—there is no fundamental limit on the number of components that can be added. One unfortunate consequence of this design was that the pipeline was quite long. This cost was considered minor compared to the benefits of the design. Figure 4.5 illustrates the modular structure of the pipeline.

In addition to the substitution domains, correlation and co-occurring relationships and all the constraints, the following characteristics were taken into consideration for the pipeline:

i. The number of new variants to be generated is specified as a command line argument to the first program in the pipeline.

ii. The first program in the pipeline generates the initial variants and does not have any input, whereas all the other programs read as input the output generated from the program preceding it in the pipeline.

iii. Since the maximum length of the mature part of the 190 fHbp sequences was 263, a maximum of 263 was used for the length of new variants.

iv. The final output is the set of variants of fHbp generated by the last program in the pipeline.

v. The amino acid substitution domains for the variable positions, the correlation and co-occurring characteristics, and the constraints were incorporated in the pipeline of programs.

vi. The constraints associated with separate positions can and did interact in complex ways; for instance, programs in the beginning of the pipeline generated the amino acids for a particular position following the constraint associated with the positive selection sites, and some of these positive selection sites may have correlated or co-occurring relationships with other positions which were not positive selection sites. When the amino acids for a particular position following the constraints associated with the correlated or co-occurring characteristics were generated by programs later in the pipeline, the amino acids associated with the positive selection sites already generated would again be included in this later program due to its correlated or co-occurring characteristics. This meant that the same position could have an interference problem due to repeated generation of amino acids for these positions. For such positions there were two alternatives.

- First, if the amino acid was already set for a particular position, then we keep that setting.
- The alternative is to allow subsequent steps to modify an already placed amino acid according to the constraints at that stage.

For our program we decided to implement the first option. The alternative option can be implemented as future work (Section 6.2).

vii. A random number generator, the Perl function *rand*, was used to generate the amino acids in positions which were not conserved. Random floating-point numbers uniformly distributed between 0 and 100 were generated [78]. The occurrence frequencies of the amino acids in each position of the original 190 sequences determined in step 6 were translated to a scale of 100% (Figure 4.6 where the scale of 100 is given to one decimal place). The scale is divided into n intervals, where n is the number of different amino acids which occur at that position. The size

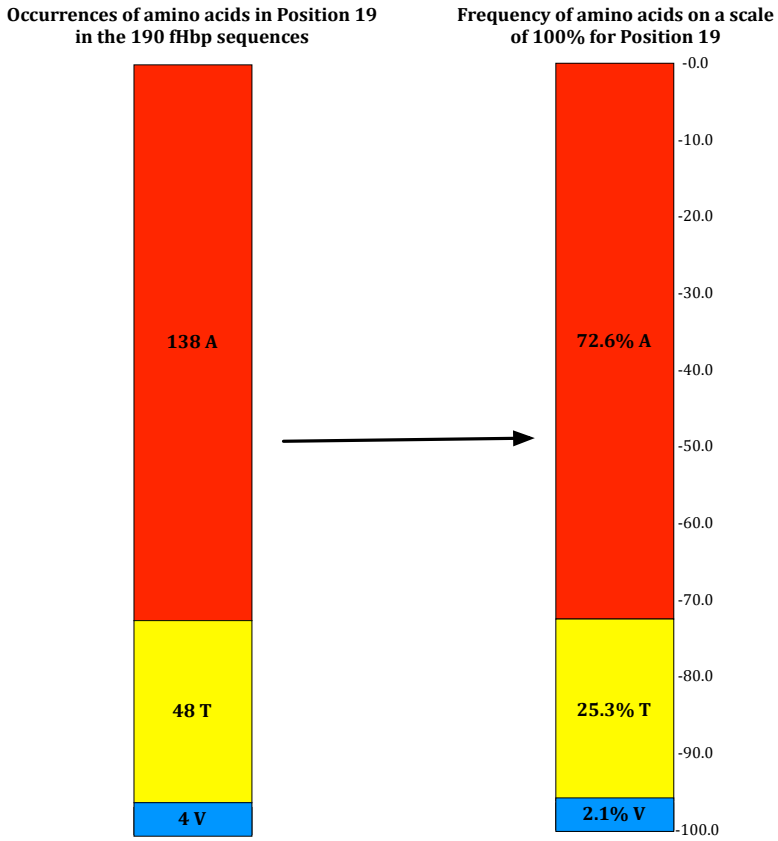


Figure 4.6: Frequency of amino acids for a particular position of the 190 fHbp sequence alignment represented on a scale of 100%

of each interval is proportional to the occurrence frequency of the corresponding amino acid at that position. By mapping the generated random number (between 0 and 100) onto this scale, an amino acid can be generated with the appropriate frequency. Figure 4.6 illustrates this concept with an example involving position 19 of the alignment. At this position, only three amino acids appeared: A, T, and V. Amino acid A appeared 72.6% of the time, T 25.3%, and V 2.1%. To produce an amino acid for this position, a random number uniformly distributed between 0 and 100 is generated as described above. If the random number was between 0 to 72.6 (not inclusive) then an 'A' was produced for that position. If the random number was between 72.6 to 97.9 (not inclusive) then a 'T' was produced. Finally, if the random number was between 97.9 to 100 (not inclusive) then a 'V' was produced.

The *rand* function was also used in setting the correlation or co-occurring relationships between the different positions in the alignment. The same random number was used to generate the amino acids for the correlated and the "fully" and "partially" co-occurring positions, so that the frequency at which the amino acids occurred remained the same for all these positions. For example, position 19 illustrated in Figure 4.6 is "fully" co-occurring with position

20, where the amino acids, C, I, and L appear. Since the positions are "fully" co-occurring the amino acids will occur in the same percentages; that is, C appearing 72.6%, I 25.3%, and L 2.1% in relation with amino acids A, T, and V respectively of position 19. The amino acids for both the positions will be generated by the same random number and in one single program of the pipeline.

As mentioned above, the Perl programs were coordinated in the pipeline using a UNIX shell script. The input for each program was read from standard input from the program preceding it in the pipeline except for the first program in the pipeline. The output of each program was generated on the standard output. This was done to facilitate inter-operability and for monitoring the operation of the pipeline.

Algorithm 4.1 Algorithm for the first program in the pipeline.

```
1  Read the number of variants to be generated as input from the command line argument
2  Set the maximum length of the variants to 263 amino acids long
3  For each new variant to be generated
4      Mark the first invariable part of the amino-terminal repetitive region, 'CSSG'
5      For the next six variable positions of the amino-terminal repetitive region which
       is comprised of glycines (G), serines (S), or 'gap' '–'
6          Use the rand function to generate a random floating point number on a scale of 0 to 100
7              Mark 'G', 'S', or '–' based on the random value
8      Mark the last invariable part of the amino-terminal repetitive region, 'GG'
9      Mark the five invariant segments which flank the variable regions V_A - V_E
10     Mark the five modular variable regions V_A - V_E as 'X'
11 Output the initial variants to standard output
```

Algorithm 4.2 Algorithm for the program which denotes the negative selection sites as 'n' and the positive selection sites as 'p'.

```
1  Read a line from standard input and store it in a two-dimensional array where index i
   represents each variant (row) and index j columns (positions in the variant)
2  For each variant (row)
3      Locate the negative selection sites
4          Mark 'n' as a symbol for the negative selection sites which are currently using 'X'
5      Locate the positive selection sites
6          Mark 'p' as a symbol for the positive selection sites which are currently using 'X'
7      Mark the remaining positions unchanged
8  Output the result to standard output
```

Algorithm 4.1, the first program in the pipeline, outputs the amino-terminal repetitive region and the invariant segments that flank the modular variable regions V_A through V_E. For the variable region in the amino-terminal repetitive segment, the program generates 'G', 'S' or 'gap' ('–') using the random number based generator explained above. For the remaining modular variable regions (V_A, V_B, V_C, V_D, and V_E), it outputs 'X'. Figure 4.7 shows portions of the output generated from this algorithm: the amino-terminal repetitive region, invariant segments 'SRFDF' and 'GEFQ', and the two modular variable regions V_A and V_B (flanked by 'SRFDF') marked by 'X'. The 8 lines

Algorithm 4.3 Algorithm for a program which determines the amino acids for positions with correlations.

```
1  Read a line from standard input and store it in a two-dimensional array where index i
   represents each variant and index j columns within the variants
2  For each variant
3      Use the rand function to generate a random floating point number on a scale of 0 to 100
4      For each position which is correlated to other position(s)
5          If the position which is correlated to other position(s) is 'X', 'n', or 'p' and if the
           random value generated corresponds to outputting the amino acid with
           correlational characteristics
6              Check the other positions with which this position is correlated
7                  If they are 'X', 'n', or 'p'
8                      Mark the amino acids that should be in all the positions
                       for the correlation to exist
9                  If the required amino acids for correlational characteristic already exists
10                     Mark the amino acid in the position which is correlated to others
11                 If some other amino acids or gap '–' exists
12                     Go back to step 3 and generate a new random number in order to
                       mark a different amino acid
13     Mark the remaining positions unchanged
14 Output the result to standard output
```

Algorithm 4.4 Algorithm for a program which determines the amino acids for positions with "fully" and "partially" co-occurring relationships.

```
1  Read a line from standard input and store it in a two-dimensional array where index i
   represents each variant and index j columns within the variants
2  For each variant
3      Use the rand function to generate a random floating point number on a scale of 0 to 100
4      For each position with "fully" or "partially" co-occurring relationship
5          Check if it is 'X', 'n', or 'p'
6              Mark the amino acids or gaps '–' allowed for this position based on the
               random value generated by the rand function
7      Mark the remaining positions unchanged
8  Output the result to standard output
```

Algorithm 4.5 Algorithm for a program which determines the amino acids for conserved positions.

```
1  Read a line from standard input and store it in a two-dimensional array where index i
   represents each variant and index j columns within the variants
2  For each variant
3      Locate the position that is supposed to be conserved
4          Check if it is 'X', 'n', or 'p'
5              Mark the amino acid intended for the conserved position
6      Mark the remaining positions unchanged
7  Output the result to standard output
```

Algorithm 4.6 Algorithm for a program which determines the amino acids for restricted variable positions with no correlation or co-occurring relationships.

```
1  Read a line from standard input and store it in a two-dimensional array where index i
   represents each variant and index j columns within the variants
2  For each variant
3      Use the rand function to generate a random floating point value on a scale of 0 to 100
4      Locate the restricted variable position
5          Check if it is 'X', 'n', or 'p'
6              Mark the amino acids according to the random value generated in step 3
7      Mark the remaining positions unchanged
8  Output the result to standard output
```

in Figure 4.7 are randomly selected from the output file of Algorithm 4.1 and shows positions 1 through 105.

Algorithm 4.2 symbolizes the negative selection sites as 'n' and the positive selections sites as 'p' only if they are already designated as 'X'. The negative and positive selection sites are derived from Step 8. It is observed that none of the negative or positive selection sites overlapped with the invariant segments and all are in the variant regions. Portions of the output generated by this algorithm are shown in Figure 4.8, where the negative selection sites are marked by 'n' and the positive selection sites by 'p'. The 8 lines and the positions 1 through 105 in this figure correspond to those in Figure 4.7.

Algorithm 4.3 outputs the amino acids for the correlated positions. It generates a random value using the *rand* function based on the concept depicted in Figure 4.6 in order to determine whether the amino acid required to establish the correlational characteristic will be output. In addition, it checks if the position correlated to other positions is designated as 'X', 'n', or 'p'. If all these conditions are satisfied, the program checks the positions to which this particular position is correlated. If these positions are either 'X', 'n', or 'p' then the amino acids for all the positions involved in the correlation are marked with the appropriate amino acids. If not, the program checks if these positions already had the required amino acids for correlational characteristic, and then the program marks the position which is correlated to these positions with the appropriate amino acid. If this condition also turns out to be false and the positions had either different amino acids or 'gap' ('–'), then the program loops back to step 3 and generates a new random value to mark the position with different amino acids. Figure 4.9 shows 8 lines (positions 34 through 130) which are randomly selected from the output of Algorithm 4.3 and illustrates correlation between the two positions, 122 and 125. This is a case where the position of interest (position 125) is correlated to a position preceding it in the alignment (position 122). In this example, whenever there is an 'L' in position 125 marked by the red arrow, there is always an 'E' in position 122 marked by the black arrow. However, the reverse is not true. Hence, position 125 is correlated to position 122 but not *vice-versa*.

Algorithm 4.4 illustrates a program which outputs amino acids for the positions with "fully" and


CSSGGG-S-SGGXXX

Figure 4.8: Portions of the output generated by Algorithm 4.2, illustrating the negative and the positive selection sites as 'n' and 'p' respectively.

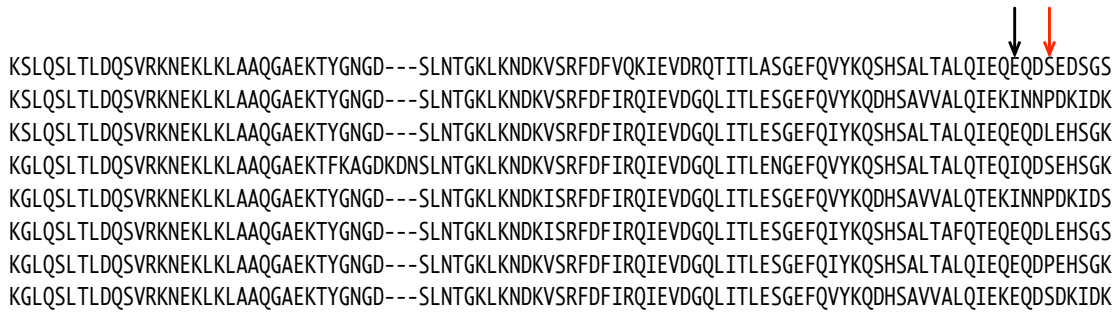


Figure 4.9: Portions of the output generated by Algorithm 4.3, illustrating correlation. Position marked by the red arrow is correlated to the position marked by the black arrow.

Figure 4.10: Portions of the output generated by Algorithm 4.4 illustrating "fully" and "partially" co-occurring relationships. The blue arrows mark the "fully" co-occurring positions and the red arrows mark the "partially" co-occurring ones.

Step 11

In this step, the set of new variants generated in the previous step was examined for duplicates and to determine if any of the new variants were duplicates of any of the original 190 fHbp sequences. To accomplish this, CD-HIT was run at 100% sequence identity on a merged file containing the 190 fHbp sequences and the set of new variants. The second task was to determine the conserved, restricted variable regions, and the amino acid substitution domains for the variable positions in the variants; for this the scripts of Appendix A were used again. The next task was to determine if the same correlations, co-occurrences and constraints determined for the existing sequences also exhibited in the variants.

Step 12

This step begins a loop involving Steps 13, 14, and the two branches of Step 15. This loop was executed for each new variant that was generated in Step 10 and studied for validity in Step 11. The loop involves predicting the tertiary structure of each new variant and then determining and recording the energy values of the structure. The last task involved validating the new variants based on their energy values; that is, if the energy values fell within the boundary defined by the existing fHbp sequences.

Step 13

This step determines the structures of each of the new variants. The same tools were used as those for the original fHbp sequences: SWISS-MODEL [13, 33, 47, 67] to derive the tertiary structures of the new variants and Swiss-PdbViewer [20, 45] for visualizing the structures. It was also recorded whether SWISS-MODEL used the same model structures for the new variants as it used for the existing fHbp sequences.

Step 14

This step involves the energy determination of the structures of the variants derived in the previous step. The same procedure and tools were used as were used for the original fHbp sequences. At first, energy minimization of all the variants was performed using Swiss-PDBViewer [20, 45], so that the same procedure was followed for both the existing sequences and predicted variants.

The next phase of this step determines and records the energy values of the structures of all the new variants using FoldX [44, 72, 73, 74], as was done for the original fHbp sequences. As each energy value was calculated, it was checked against the valid boundary energy value determined in Step 9 from the energy values of the 190 fHbp sequences.

Step 15

In this step, the energy values determined in the previous step were examined and the disposition of the new variant was determined as follows:

- ***Step 15 i***

  If the energy value of the variant was within the valid boundary for energy, then the variant was recorded as a valid variant.

- ***Step 15 ii***

  If the energy value of the variant was above the valid boundary, then the variant was discarded as an invalid variant.

After Steps 12 to 15 were executed for all the variants, the result was a set of valid new variants. These valid new variants of fHbp were the ones that nature might "allow" but which have not as yet appeared. The end result of this research is illustrated by the parallelogram highlighted in yellow in the methodology flow chart (Figure 4.1).

Chapter 5

Results

This chapter presents, discusses, and analyzes the results generated in this research.

5.1 Identifying the amino-terminal repetitive region and the five modular variable regions

This section discusses the results generated from Steps 1 to 4 of the Methodology (Figure 4.1).

In total, 270 sequences of fHbp were downloaded from UniProtKB out of which the duplicates and fragments were discarded using the CD-HIT algorithm. This reduced the number of sequences to 200. The 200 sequences were then aligned using ClustalW. Further study of the multiple sequence alignment revealed that there were ten sequences which did not conform to the pattern for fHbp identified by Beernink et al. [34] (Section 2.6). The 10 sequences were discarded leaving a total of 190 fHbp sequences.

5.2 Identifying the mature protein part

This section discusses the results generated from Step 5 of the Methodology (Figure 4.1).

The first task was to determine the consensus sequence pattern for fHbp identified by Beernink et al. [34]. EMBOSS *cons* algorithm was used to generate the consensus sequence illustrated in Figure 5.1.

The portion up to and including the amino terminal repetitive segment (designated as the "N-term" in Figure 2.2) of the consensus sequence (Figure 5.1) was then input into the SignalP 3.0 Server [17]—as discussed in Step 5 of Section 4.2—to determine the signal and mature protein portion of the sequence. The black arrow in Figure 5.1 illustrates the end of the region up to and including the amino terminal repetitive segment in the consensus sequence and comprises of 37 peptides.

Figure 5.2 illustrates the output generated by SignalP 3.0 server using a Neural Network (SignalP-NN). The SignalP-NN output is comprised of scores, *C*, *S*, *Y*, *S-mean*, and *D*. The *S* score reports the signal peptide prediction for each amino acid, with high scores indicating that the

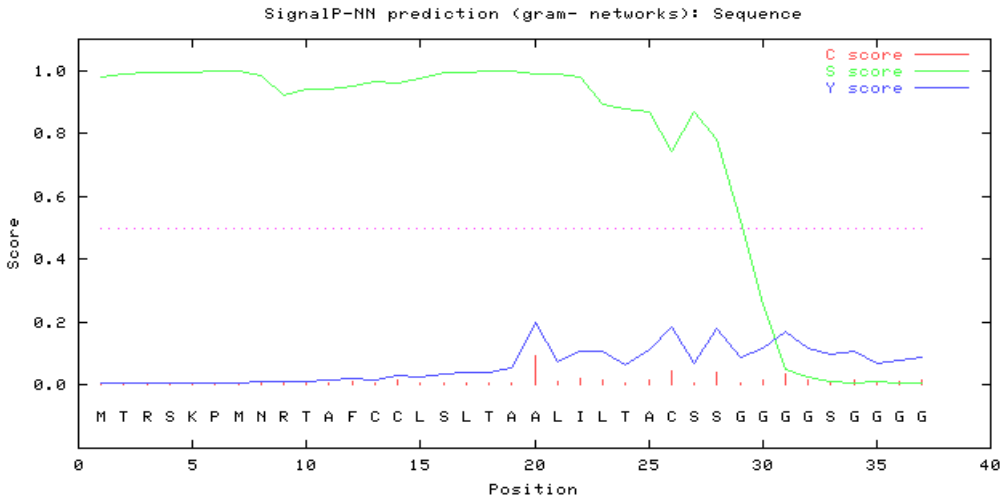
Figure 5.1: The consensus sequence of the 190 fHbp sequences. The black arrow marks the end of the portion up to and including the amino terminal repetitive segment.

amino acid is part of the signal peptide, and low scores indicating that the amino acid is part of the mature protein part [16]. The tabular output in Figure 5.2 shows that the amino acids starting from position 1 have a high *S-score*, with the values decreasing somewhat over positions 23 to 28. There is a sharp decrease in score from positions 28 to 31, with the score continuing to be low for the remaining amino acids. The graph in Figure 5.2 illustrates the same output with the S score highlighted in light green colour. The C score is the "cleavage site" score, which marks the end of the signal part and the beginning of the mature protein part [16, 35, 63]. The Y score is a derivative of the C score combined with the S score, resulting in a better cleavage site prediction than the raw C score alone. This is because multiple high-peaking C scores can be found in a sequence, where only one is the actual true cleavage site [16, 35, 63]. From the graph of Figure 5.2, it can be seen that there are multiple cleavage sites marked by both the red bars (C scores) and the blue line (Y score) with the residue cysteine (C) of 'CSSG' region marked as one of the cleavage sites. However, the highest C and Y scores are in position 20, indicating a possibility of the start of the matured protein part. The *S-mean* score is the average of the S score, ranging from the N-terminal amino acid to the amino acid assigned with the highest *Y-max* score; that is, the *S-mean* score is calculated for the length of the predicted signal peptide. The *D-mean* score is the average of the *S-mean* and *Y-max* score, which also determines the signal peptide region [16, 35, 63]. In Figure 5.2, the values of the *S-mean* and *D-mean* score indicates that the signal peptide region is from position 1 to 19 with the amino acid in position 20 serving as the first amino acid of the mature protein part.

Figure 5.3 illustrates the output generated by the SignalP 3.0 Server using the Hidden Markov Models (SignalP-HMM). The Hidden Markov Model calculates the probability of whether the submitted sequence contains a signal peptide or not; this is depicted by the S score [16]. Figure 5.3 shows the probability of being a signal peptide, illustrated by the S score being 1 for positions 1 to 25, and decreasing sharply at position 26 which marks the amino acid cysteine (C) of the 'CSSG' region. From there onwards, the probability decreases, leading to a score of 0 from position 33. The

```
>Sequence                 length = 37

# pos  aa     C        S        Y
    1   M    0.006    0.979    0.006
    2   T    0.006    0.988    0.005
    3   R    0.006    0.995    0.005
    4   S    0.006    0.995    0.005
    5   K    0.006    0.996    0.006
    6   P    0.006    0.998    0.007
    7   M    0.007    0.998    0.008
    8   N    0.007    0.986    0.011
    9   R    0.007    0.920    0.013
   10   T    0.008    0.939    0.014
   11   A    0.006    0.942    0.015
   12   F    0.014    0.952    0.023
   13   C    0.006    0.964    0.017
   14   C    0.015    0.961    0.033
   15   L    0.007    0.976    0.028
   16   S    0.007    0.994    0.034
   17   L    0.007    0.992    0.042
   18   T    0.006    0.999    0.042
   19   A    0.009    0.999    0.056
   20   A    0.094    0.987    0.199
   21   L    0.012    0.990    0.075
   22   I    0.021    0.982    0.106
   23   L    0.019    0.892    0.107
   24   T    0.007    0.878    0.066
   25   A    0.018    0.867    0.113
   26   C    0.046    0.746    0.186
   27   S    0.006    0.870    0.070
   28   S    0.039    0.780    0.180
   29   G    0.009    0.524    0.087
   30   G    0.016    0.255    0.119
   31   G    0.035    0.052    0.173
   32   G    0.017    0.027    0.116
   33   S    0.013    0.012    0.099
   34   G    0.018    0.009    0.109
   35   G    0.008    0.012    0.071
   36   G    0.012    0.007    0.080
   37   G    0.016    0.008    0.089
```



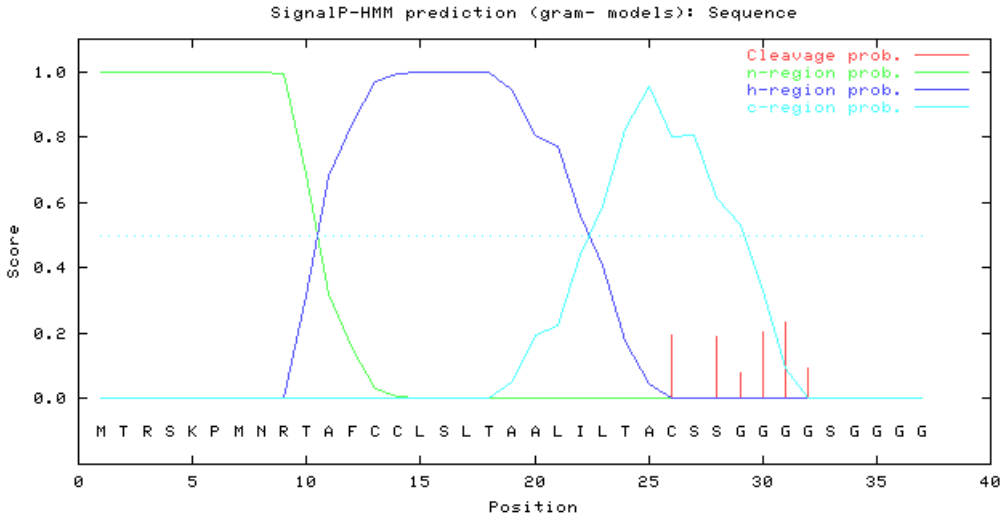

```
>Sequence                 length = 37
# Measure  Position  Value  Cutoff  signal peptide?
  max. C     20       0.094   0.52   NO
  max. Y     20       0.199   0.33   NO
  max. S     18       0.999   0.92   YES
  mean S      1-19    0.978   0.49   YES
       D      1-19    0.588   0.44   YES
# Most likely cleavage site between pos. 19 and 20: LTA-AL
```

Figure 5.2: SignalP-NN output. This graphical output from SignalP 3.0 Server (Neural Network) was based on a combination of several artificial neural networks.

SignalP-HMM result:

>Sequence

# pos	aa	C	S	n-reg	h-reg	c-reg
1	M	0.000	1.000	1.000	0.000	0.000
2	T	0.000	1.000	1.000	0.000	0.000
3	R	0.000	1.000	1.000	0.000	0.000
4	S	0.000	1.000	1.000	0.000	0.000
5	K	0.000	1.000	1.000	0.000	0.000
6	P	0.000	1.000	1.000	0.000	0.000
7	M	0.000	1.000	1.000	0.000	0.000
8	N	0.000	1.000	1.000	0.000	0.000
9	R	0.000	1.000	0.996	0.004	0.000
10	T	0.000	1.000	0.685	0.315	0.000
11	A	0.000	1.000	0.321	0.679	0.000
12	F	0.000	1.000	0.152	0.847	0.000
13	C	0.000	1.000	0.029	0.971	0.000
14	C	0.000	1.000	0.006	0.994	0.000
15	L	0.000	1.000	0.001	0.999	0.000
16	S	0.000	1.000	0.000	0.999	0.000
17	L	0.000	1.000	0.000	1.000	0.000
18	T	0.000	1.000	0.000	0.998	0.002
19	A	0.000	1.000	0.000	0.948	0.052
20	A	0.000	1.000	0.000	0.806	0.194
21	L	0.000	1.000	0.000	0.774	0.225
22	I	0.000	1.000	0.000	0.559	0.441
23	L	0.000	1.000	0.000	0.409	0.591
24	T	0.000	1.000	0.000	0.174	0.826
25	A	0.000	1.000	0.000	0.046	0.954
26	C	0.195	0.805	0.000	0.002	0.802
27	S	0.000	0.805	0.000	0.000	0.805
28	S	0.192	0.612	0.000	0.000	0.612
29	G	0.079	0.533	0.000	0.000	0.533
30	G	0.205	0.328	0.000	0.000	0.328
31	G	0.232	0.096	0.000	0.000	0.096
32	G	0.096	0.000	0.000	0.000	0.000
33	S	0.000	0.000	0.000	0.000	0.000
34	G	0.000	0.000	0.000	0.000	0.000
35	G	0.000	0.000	0.000	0.000	0.000
36	G	0.000	0.000	0.000	0.000	0.000
37	G	0.000	0.000	0.000	0.000	0.000



>Sequence
Prediction: Signal peptide
Signal peptide probability: 1.000
Max cleavage site probability: 0.232 between pos. 30 and 31

Figure 5.3: SignalP-HMM output. This output was based on the Hidden Markov Models which calculated the probability of whether the submitted sequence contained a signal peptide or not.

C score is the "cleavage site" score similar to the SignalP-NN output (Figure 5.2). The first cleavage site in Figure 5.3 is position 26 (C of 'CSSG'). However the maximum cleavage site probability is shown between the positions 30 and 31. The same results are illustrated in the graph of Figure 5.3, marked by the red bars. The *n-reg*, *h-reg*, and *c-reg* in Figure 5.3 represents the predicted positively charged amino-terminal region, the hydrophobic region, and the polar carboxy-terminal region, respectively, of the signal protein part [77]. From the data and the graph of Figure 5.3, it can be seen that the *n-region* of the signal peptide is from position 1 to 8 with the scores gradually decreasing from position 9. The *h-region* score increases from position 9 and starts decreasing after position 17. The *c-region* score starts increasing from position 18 and starts dropping after position 25.

From the overall analysis of the outputs of both the SignalP-NN and SignalP-HMM and from the findings of Masignani et al. [55] (mentioned in Step 5 of Section 4.2), it can be concluded that the mature protein part of the fHbp sequence starts from the cysteine (C) of the 'CSSG' region.

Appendix C (Figures C.1 to C.6) show parts of the multiple sequence alignment of the 190 fHbp sequences with the signal protein part, the amino-terminal repetitive region, the modular variable regions, and the invariant regions highlighted with different colours.

After the detection of the signal protein part of the sequences, the task was to study them. It was seen that among the 190 sequences, 50 of them had the signal protein part and 140 did not. Most of the positions in the signal part were conserved. There was no correlation or co-occurring relationships observed amongst the variable positions within the signal part or with the positions of the mature part. The next task was to determine if the presence of the signal part made any difference to the predicted tertiary structures of the sequences. For this purpose around 20 sequences which had the signal part were randomly selected. Two things were observed during this investigation. First, it was seen that for some of the selected sequences with the signal part, the presence of the signal peptides had impact on the PDB [24] template SWISS-MODEL was using to derive the tertiary structures of the sequences. With the signal protein part, for some of the selected sequences, SWISS-MODEL used 2W80H (one of the chains of the fHbp variant 2W80 in PDB) as its template. With the removal of the signal part and using only the mature part of the sequence, SWISS-MODEL used 2W80C (which is referred to as the unique factor H binding protein chain of 2W80 in PDB) as the template. The reverse was also true, when 2W80C was used as the template in the presence of the signal part and 2W80H when the signal part was removed. Among such cases, the second observation was related to the percentage identity with the template used to derive the tertiary structure. It was seen that the presence of the signal parts in the sequences affected the percentage identity with the templates used to predict the structures. Both these observations were significant because these predicted tertiary structures would later be used to determine the energy values of the sequences and to set the valid range of energy the

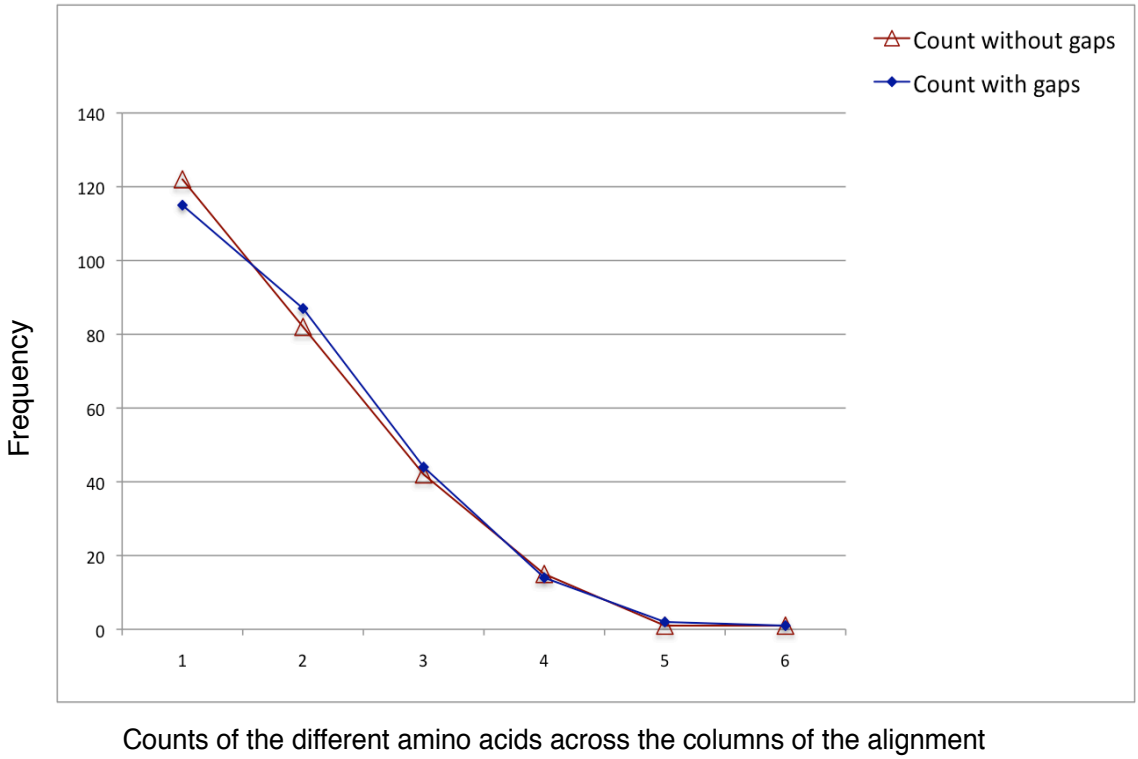


Figure 5.4: Graph showing the frequency of the different amino acids occurring in the positions of the multiple sequence alignment of the 190 fHbp sequences. The blue curve was based on data which included 'gap' in the count for each position. The red curve was based on counts which ignored the presence of 'gap'. Only positions of the mature part of the protein sequence were considered.

new variants can have. In addition, as mentioned earlier, the presence of the signal part was not consistent with all the sequences. Hence, it was decided to discard the signal protein part of the fHbp sequences. This reduced the length of the longest fHbp sequence from 291 amino acids to 263.

5.3 Identifying the conserved, restricted variable, and unrestricted variable regions

The results described in this section were produced by the Scripts in Appendix A as part of Step 6 of the Methodology (Figure 4.1) in order to identify the conserved, restricted variable, and unrestricted variable regions, and the amino acid substitution domains for the variable regions in the mature protein part of the sequences. The results generated are given in tabular formats in Appendix B (Tables B.1 to B.5).

In Tables B.1 and B.2 the columns headed "190 fHbp sequences" depict the substitution domains

for each of the restricted variable positions; that is, which amino acids were observed in each of the positions of the restricted variable regions and in addition the frequencies of the occurrences of each amino acid. Hence, the results in the tables presents occurrence profiles for each position along the alignment of the 190 fHbp sequences. The first table generated using Script A.1 excludes the count of the 'gap' symbols, and the second one generated using Script A.2 includes the count of the 'gap' symbols.

The column labeled "190 fHbp sequences" in Table B.3 illustrates which positions in the alignment had at least one 'gap' symbol and is generated using Script A.3.

Scripts A.4 and A.5, determined the frequency distribution for the counts of the different amino acids across the columns of the alignment, both excluding and including the 'gap'. Tables B.4 and B.5 (Appendix B) give those results for the 190 fHbp sequences. They are illustrated graphically in Figure 5.4. These results illustrate that the amino acid substitutions were limited; that is, variation in the sequence alignment was restricted. Out of the 263 positions of the alignment, almost half of them had only one type of amino acid (122 positions with the 'gap' symbol excluded and 115 positions with the 'gap' symbol included), meaning that these positions were conserved. Among these conserved positions, 24 positions belonged to the invariant positions in the amino-terminal repetitive region and the five invariant segments that delimited the variable regions V_A through V_E (illustrated as black bars in Figure 2.2). Also the maximum number of different amino acids occurring at any one position was 6 (both including and excluding the 'gap' symbol) and that was only for one position. Only 16 positions had 4 or 5 different amino acids, further supporting the conjecture that variation in the alignment was highly restricted. Hence, it was deduced that the mature part of the 190 fHbp sequence alignment had conserved and restricted variable segments but no unrestricted variable regions.

5.4 Identifying correlation and co-occurrence between positions with restricted variations

The results discussed here were generated as part of Step 7 of our Methodology (Figure 4.1). It was observed in the multiple sequence alignment of the 190 fHbp sequences that mutations in one position were correlated or co-occurring (both "partially" and "fully") with sometimes one position and sometimes groups of positions in different variable regions. However, it was seen that there were more "partially" and "fully" co-occurring regions than there were correlated ones. One reason could be because correlations were examined visually and some could have been missed. It was out of the scope of this thesis to use an automated tool for the examination. However, in future, this can be done using a statistical or a computational tool for better accuracy, as mentioned in Section 6.2. Figures 5.5 and 5.6 are portions of the multiple sequence alignment of the 190 fHbp sequences

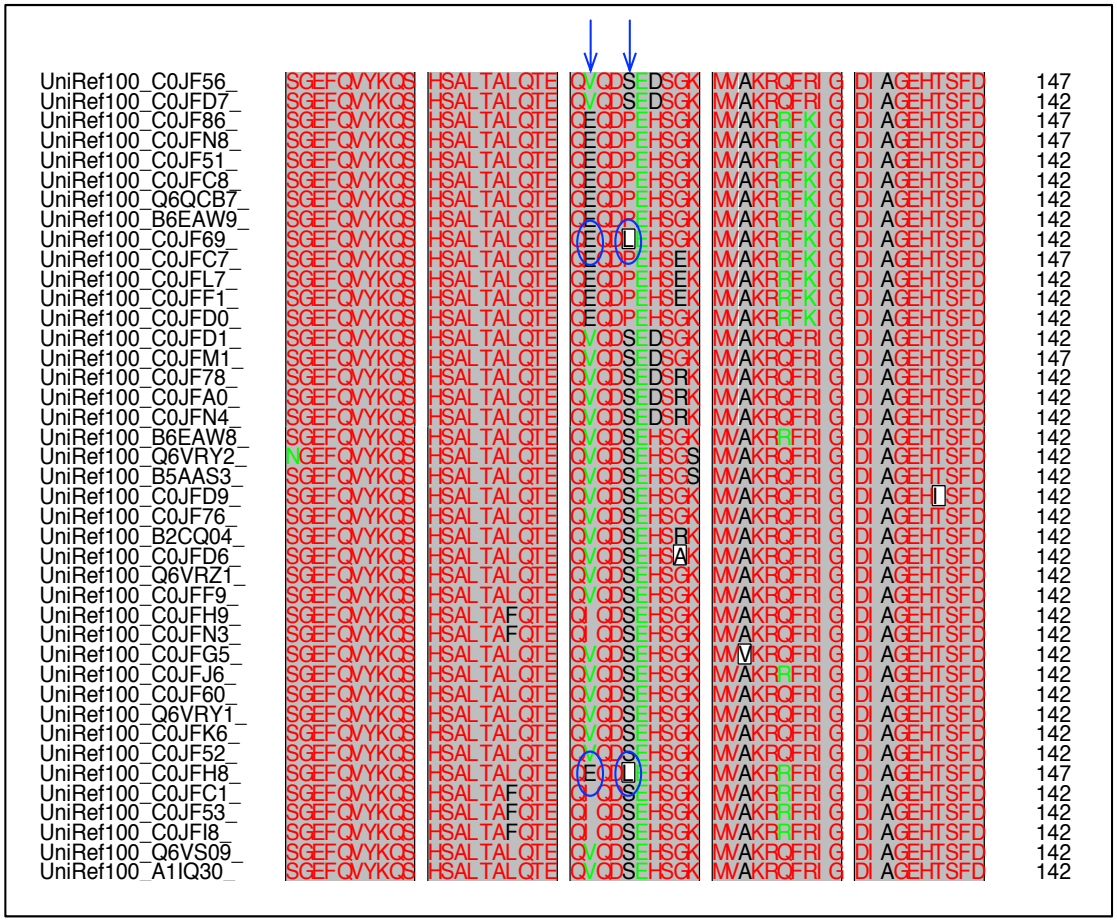


Figure 5.5: An example of correlation between amino acids in different positions.
This portion of the multiple sequence alignment was taken from the alignment of the 190 fHbp sequences. The two positions marked by the blue arrows had a correlated relationship. The correlated amino acids are marked by the blue circles. Whenever there was a leucine (L) in the position marked by the second arrow, there was always a glutamic acid (E) in the other position. However, the reverse was not true. For glutamic acid (E) in the first marked position, there were other amino acids apart from leucine (L) in the second marked position (for instance, P). Hence, if the second marked position was 'L', then the first marked position would have to be an 'E', thus establishing a correlated relationship.

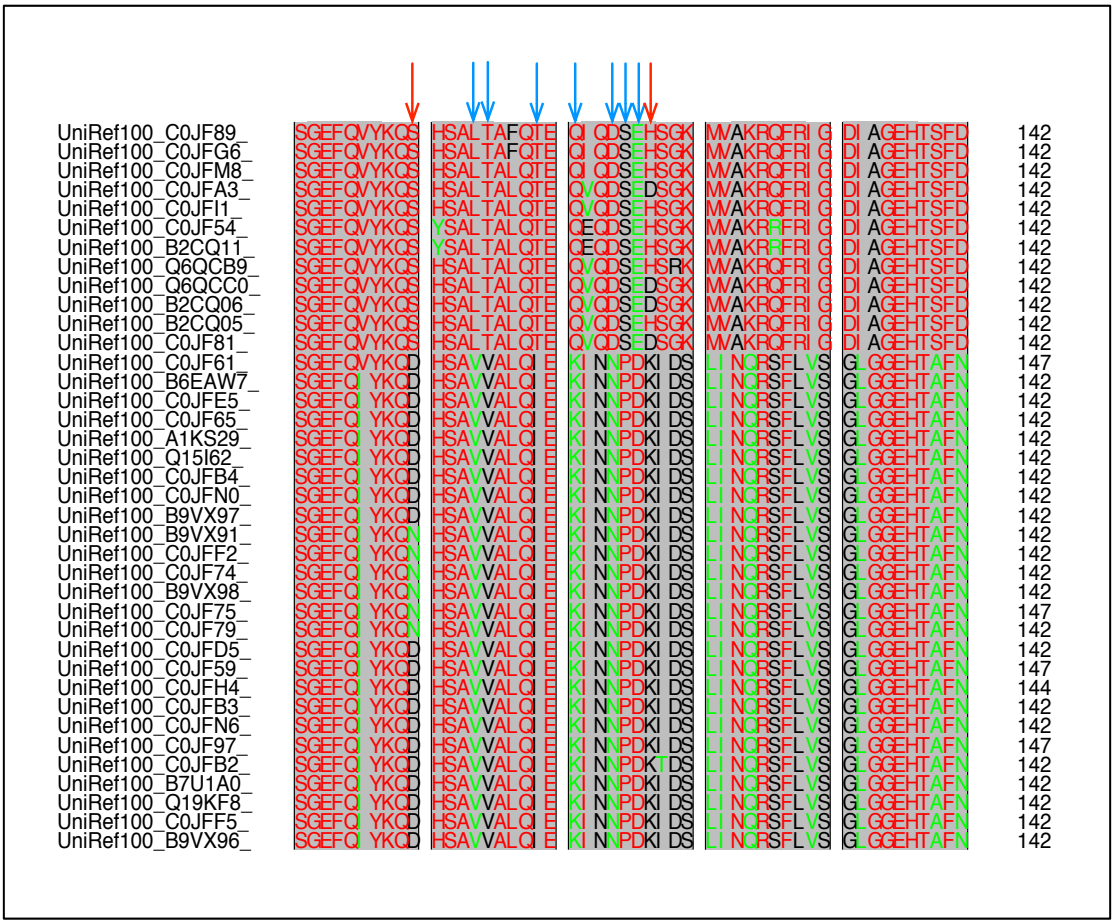


Figure 5.6: An example of co-occurring relationships between amino acids in different positions. This portion of the multiple sequence alignment was taken from the alignment of the 190 fHbp sequences. Positions marked by the blue arrows are "fully" co-occurring and those marked with red arrows are "partially" co-occurring. It can be seen that the marked positions are either always 'S', 'L', 'T', 'T', 'Q', 'D', 'S', 'E', 'H/D' respectively, or else they are 'D/N', 'V', 'V', 'I', 'K', 'N', 'P', 'D', 'K' respectively, hence establishing a co-occurring relationship amongst them.

illustrating correlated and both kinds of co-occurrence relationships among different positions in the alignment.

In Figure 5.5 the blue arrows mark two positions, 122 and 125, involved in a correlated relationship, the only correlation found in the 190 fHbp sequences. Whenever there was a leucine (L) in position 125 (marked by the second blue arrow), there was always a glutamic acid (E) in position 122 (marked by the first blue arrow). However, the reverse was not true; that is, whenever there was a glutamic acid (E) in position 122, there were other amino acids in addition to leucine (L) in position 125. This correlation can be denoted as L→E.

In Figure 5.6 the blue arrows mark "fully" co-occurring positions, and the red arrows "partially" co-occurring ones. Both the blue and red arrow marked positions have co-occurring relationships amongst themselves. Starting from the left hand side, the columns marked are either always 'S', 'L', 'T', 'T', 'Q', 'D', 'S', 'E', 'H/D' respectively, or else they are 'D/N', 'V', 'V', 'I', 'K', 'N', 'P', 'D', 'K' respectively, hence establishing a co-occurring relationship amongst them.

Table 5.1 lists the instances of co-occurrences in the 190 fHbp sequences. As shown in the table, there were 7 instances of co-occurring relationships comprising both "fully" co-occurring and "partially" co-occurring positions, involving a total of 85 positions. The instances are highlighted in alternating red and white colour for clarity. Each block (either red or white) composed of several "fully" and "partially" co-occurring positions (rows) represents each instance of co-occurrence. The percentages in brackets are the percentage occurrence of the amino acids in each position. The first instance (highlighted with red) of co-occurrence was between positions 37, 42, 43, 45, 46, 47, 49, 50, 52, and 54 where 43, 45, 49, 50, and 52 belonged to the "fully" co-occurring group and 37, 42, 46, 47, and 54 to the "partially" co-occurring group. In this instance, the percentage distribution between the amino acids "fully" co-occurring with each other is 72.1% and 27.9%; that is, in these positions, two different amino acids occur and in fixed proportion. In the "partially" co-occurring positions even though two or more different amino acids occur, the percentage distribution is maintained. For instance, in position 37, amino acids Q (which occurs 70%) and R (which occurs 2.1%) constitute a 72.1% portion of the percentage distribution. In other words, whenever there is Q in position 43, there is always either Q or R in position 37, and whenever there is D in position 43 there is amino acid K in position 37, thus establishing a co-occurring relationship among them.

Table 5.1: "Fully" and "partially" co-occurring positions and their corresponding amino acids in the 190 fHbp sequences.

"Fully" co-occurring positions	Corresponding amino acids	"Partially" co-occurring positions	Corresponding amino acids
43	Q (72.1%), D (27.9%)	37	Q (70%), R (2.1%), K (27.9%)
45	V (72.1%), I (27.9%)	42	D (71.6%), N (0.5%), E(27.9%)
49	E (72.1%), G (27.9%)	46	R (72.1%), P (18.4%), S (9.5%)
50	K (72.1%), T (27.9%)	47	K (71.6%), R (0.5%), Q (27.9%)
52	K (72.1%), T (27.9%)	54	A(70%), S(30%)
62	Y (73.7%), F (26.3%)	64	N (73.7%), A (22.1%), V (4.2%)
63	G (73.7%), K (26.3%)		
67	– (73.7%), K (26.3%)		
68	– (73.7%), D (26.3%)		
69	– (73.7%), N (26.3%)		
87	I (77.9%), V (22.1%)	88	R(75.8%), H (2.1%), Q (22.1%)
89	Q (77.9%), K (22.1%)		
96	L (74.7%), T (25.3%)		
100	E (74.7%), A (25.3%)		
114	L (55.8%), V (44.2%)	110	S (55.8%), D (35.8%), N (7.9%), G (0.5%)
115	T (55.8%), V (44.2%)	127	H (38.4%), D (17.4%), K (44.2%)
121	Q (55.8%), K (44.2%)	128	S (55.8%), I (40.5%), T (3.7%)
123	Q (55.8%), N (44.2%)	129	G (49.5%), E (3.2%), R (2.6%), A (0.5%), D (44.2%)
124	D (55.8%), N (44.2%)	133	A (55.3%), V (0.5%), N (44.3%)
126	E (55.8%), D (44.2%)	134	K (55.8%), Q (43.7%), R (0.5%)
131	M (55.8%), L (44.2%)	136	Q (41.1%), R (14.7%), S (44.2%)
132	V (55.8%), I (44.2%)	138	R (47.4%), K (8.4%), L (44.2%)
139	I (55.8%), V (44.2%)	141	D (55.8%), G (43.7%), S (0.5%)
140	G (55.8%), S (44.2%)	143	A (52.6%), V (3.2%), G (44.2%)
142	I (55.8%), L (44.2%)	150	D (53.7%), G (2.1%), N (44.2%)
148	S (55.8%), A (44.2%)	154	K (28.9%), E (25.9%), D (0.5%), G (0.5%), – (44.2%)
151	K (55.8%), Q (44.2%)	157	R (26.8%), M (14.8%), S (14.2%), K (44.2%)
159	T (55.8%), E (44.2%)		
161	R (55.8%), H (44.2%)		
163	T (55.8%), K (44.2%)		
170	A (79.5%), T (0.5%), P (20%)	171	G (75.8%), S (3.7%), R (0.5%), N (20%)
181	A (79.5%), V (0.5%), T (20%)	173	K (79%), E (0.5%), R (20.5%)
188	K (79.5%), G (0.5%), R (20%)	175	T (72.1%), I (7.9%), H (20%)
		182	A (76.9%), N (2.6%), S (0.5%), K (17.4%), N (2.6%), T (1%)
194	S (55.3%), T (44.7%)	205	Y (29.5%), D (25.3%), N (0.5%), E (44.7%)
206	I (55.3%), L (44.7%)	208	P (54.8%), Q (0.5%), A (44.7%)
217	S (55.3%), L (44.7%)	212	H (27.9%), R (27.4%), S (44.7%)
220	V (55.3%), T (44.7%)	219	S (54.8%), F (0.5%), D (44.7%)
229	S (55.3%), T (44.7%)	221	L (55.3%), H (0.5%), R (44.2%)

Continued on next page

"Fully" co-occurring positions	Corresponding amino acids	"Partially" co-occurring positions	Corresponding amino acids
231	S (55.3%), H (44.7%)	223	N (55.3%), D (0.5%), G (44.2%)
233	G (55.3%), A (44.7%)	224	Q (54.8%), H (0.5%), S (30%), G (14.7%)
234	I (55.3%), L (44.7%)	225	D (33.7%), A (21.1%), N (0.5%), E (44.7%)
247	E (55.3%), T (44.7%)	237	G (49.5%), E (5.3%), R (0.5%), D (44.7%)
250	T (55.3%), I (44.7%)	238	Q (28.4%), K (26.4%), E (0.5%), R (44.7%)
252	N (55.3%), E (44.7%)	251	A (33.7%), G (21.6%), R (23.1%), V (21.6%)
253	G (55.3%), K (44.7%)		
254	I (55.3%), V (44.7%)		
256	H (55.3%), E (44.7%)		
259	L (55.3%), I (44.7%)		
261	A (55.3%), G (44.7%)		

5.5 Characterizing constraints of the restricted variable positions

This section presents the results generated in Step 8 of the Methodology (Figure 4.1) and discusses constraints characterized for positions with restricted variations.

The first task was to determine the negative selection sites in the multiple sequence alignment of the 190 fHbp sequences. Section 2.6 discusses the binding of fH and fHbp and the regions in fHbp which have high affinity for interactions with fH [70]. fHbp has an extended recognition site for fH across its entire surface. The residues along the regions of both fHbp and fH that are involved in their binding have been identified. There are 21 positions. Figure 2.1 illustrates those positions for the UniProtKB sequence Q9JXV4, one of the set of 190 fHbp sequences. The next task was to use this information in determining the negative selection sites in the alignment of 190 fHbp sequences. An intuitive assumption is that nature would try to conserve the binding regions of fHbp, and so these sites would be less likely to change with the high possibility that they would be under negative selection pressure.

Table 5.2 lists the positions in the UniProtKB sequence Q9JXV4 and the corresponding positions in the alignment of the 190 fHbp sequences. The table also gives the amino acids and their frequency of occurences at each position and illustrates to which restricted variable regions–V_A through V_E (Figure 2.2)– the binding sites belong. Position 103 of Q9JXV4 corresponds to position 43 of the alignment because the sequence Q9JXV4 includes the signal portion of the protein. The names of each of the single-letter data-base codes for the amino acids used in this table can be found in Table D.1 (Appendix D). It can be seen from the table that the binding sites are distributed across

the restricted variable regions: V_A, V_C, and V_E. None of the binding sites belong to the regions V_B or V_D and also does not overlap with the invariant segments which flanks the variable regions.

Table 5.2: Positions in fHbp that bind with fH and their corresponding variable regions.

Q9JXV4		**fHbp alignment**		**Variable Region**
Position	Amino acids	Amino acids	Position	
103	Q	137Q 53D	43	V_A
106	R	137R 35P 18S	46	V_A
107	K	136K 53Q 1R	47	V_A
108	N	187N 3K	48	V_A
180	Q	106Q 84N	123	V_C
181	D	106D 84N	124	V_C
183	E	106E 84D	126	V_C
184	H	84K 73H 33D	127	V_C
185	S	106S 77I 7T	128	V_C
191	K	106K 83Q 1R	134	V_C
193	Q	84S 78Q 28R	136	V_C
195	R	90R 84L 16K	138	V_C
262	D	85E 56Y 48D 1N	205	V_E
264	K	189K 1E	207	V_E

Continued on next page

Table 5.2 – continued from previous page

Q9JXV4		fHbp alignment		Variable Region
Position	Amino acids	Amino acids	Position	
266	D	190D	209	V_E
272	V	190V	215	V_E
274	S	105S 85L	217	V_E
283	E	190E	226	V_E
286	S	105S 85T	229	V_E
304	E	105E 85T	247	V_E
306	K	133K 57E	249	V_E

The task was now to find the positive selection sites in the alignment of the 190 fHbp sequences plus verify the negative selection sites already listed in Table 5.2. As mentioned in Step 8 of the Methodology section, determining the positive and negative selection sites required the nucleotide sequences, which were downloaded from the UniProtKB for all the 190 fHbp protein sequences. The Kumar method was used in the MEGA4 [22] tool in order to determine the K_A/K_S ratio. The K_A/K_S ratio was first found for the overall sequences and then the codon-based positive and negative selection sites using the sliding window approach were found. For this approach, the window width was 18 base pairs (6 amino acids) and the step length was 3 base pairs (1 amino acid). In the research paper where the codon-based detection of the positive and negative selection sites using the siding window approach was described [81], it was found that the window width was always 6 times the step length. Hence, in keeping consistency, for this thesis the step length was taken as 3 base pairs or 1 amino acid (the smallest base pair that could have been taken) and the window width being 6 times more was taken as 18 base pairs. Also, since the 190 fHbp sequences were all closely related, K_S was estimated from the whole sequence alignment as mentioned in Step 8 earlier. This was done because often K_S becomes 0 for closely-related sequences and/or when the window size was small [81]. For determining the K_A/K_S ratio, the nucleotide sequences corresponding to the 190 fHbp protein sequences were downloaded from the UniProtKB database [27].

The results of the Kumar Method in the MEGA4 [22] tool were as follows.

$$K_A = 0.144$$
$$K_S = 0.366$$
$$K_A/K_S = 0.411$$

These results were consistent with the finding of Brehony et al. [37] who mentioned that the fHbp locus had an average K_A/K_S ratio of 0.35, indicating an overall level of purifying selection (or

negative selection) against amino acid changes.

The focus now moves onto the codon-based detection of the positive and negative selection sites of the 190 fHbp sequences using the sliding window approach. Table B.6 in Appendix B illustrates the findings of the K_A/K_S ratio. The first column of the table shows the windows, that is, the range of the positions in the nucleotide sequences considered for the calculation. The second column shows the corresponding amino acid positions in the 190 fHbp sequence alignment. The third column gives the value of the ratio of the number of nonsynonymous substitutions per nonsynonymous site (K_A) for the particular nucleotide positions depicted in column one. The last column gives the corresponding K_A/K_S ratio. As mentioned above, K_S was determined to be 0.366.

Regions where the K_A/K_S ratio was more than 0.9 were considered to be under positive selection pressure. This was done because the K_A/K_S ratio is based on a window of 18 nucleotides, and negative selection pressure on any one of the nucleotides may affect and reduce the ratio for the entire window. These regions are highlighted with red colour in Table B.6 of Appendix B.

The K_A/K_S ratios in Table B.6 were then used to verify that each fH binding position given in Table 5.2 is under negative selection pressure. Among the 21 positions of Table 5.2, 17 positions were confirmed to be under negative selection pressure. Table 5.3 illustrates one such position. The remaining 4 among the 21 positions had conflicting K_A/K_S ratios, which are illustrated in Tables 5.4 through 5.7. Tables 5.3 through 5.7 are all portions of Table B.6.

Analyzing Amino Acid Position 43

This was the 1st position in the fHbp sequence that takes part in the binding with fH. Table 5.3 illustrates the values of the ratio of K_A/K_S for those windows in which position 43 was also a part. From the values of K_A/K_S ratio shown in Table 5.3, it was seen that all the values were less than 1, meaning that they were under negative selection pressure.

Table 5.3: Analysis of the K_A/K_S ratio using the sliding window approach for Position 43.

Window of the nucleotide positions	**Corresponding amino acid positions**	K_A	K_A/K_S
112-129	38-43	0.026	0.076
115-132	39-44	0.026	0.076
118-135	40-45	0.025	0.073
121-138	41-46	0.006	0.017
124-141	42-47	0.006	0.017
127-144	43-48	0.258	0.752

Analyzing Amino Acid Position 134

This was the 10th position that takes part in the binding with fH. Table 5.4 illustrates the values of the ratio of K_A/K_S for those windows containing position 134. All the values of the K_A/K_S ratio were less than 1, except for the last window (positions 134-139). This exception could have been due to the codon at position 139 being under positive selection pressure. This can be better understood in the analysis of the amino acid positions 136 and 138 in the next subsections.

Table 5.4: Analysis of the K_A/K_S ratio using the sliding window approach for Position 134. Positions where the K_A/K_S ratios are more than 0.9 are highlighted with red colour.

Window of the nucleotide positions	**Corresponding amino acid positions**	K_A	K_A/K_S
385-402	129-134	0.014	0.041
388-405	130-135	0.002	0.006
391-408	131-136	0.022	0.064
394-411	132-137	0.024	0.070
397-414	133-138	0.044	0.128
400-417	134-139	0.388	1.131

Analyzing Amino Acid Position 136

This was the 11th position that takes part in the binding with fH. Table 5.5 illustrates the values of the ratio of K_A/K_S for those windows containing position 136. From the values of the K_A/K_S ratio, it was seen that the last three windows (positions 134-139, 135-140, and 136-141) had values of more than 1. These exceptions could have been due to some codon starting from amino acid position 139 which might be under positive selection pressure. This is further proved in the next subsection discussing the analysis of the amino acid position 138.

For the remaining windows which included position 136, the K_A/K_S ratio was less than 1.

Table 5.5: Analysis of the K_A/K_S ratio using the sliding window approach for Position 136. Positions where the K_A/K_S ratios are more than 0.9 are highlighted with red colour.

Window of the nucleotide positions	**Corresponding amino acid positions**	K_A	K_A/K_S
391-408	131-136	0.022	0.064
394-411	132-137	0.024	0.070
397-414	133-138	0.044	0.128
400-417	134-139	0.388	1.131
403-420	135-140	0.434	1.265
406-423	136-141	0.606	1.767

Analyzing Amino Acid Position 138

This was the 12th position that takes part in the binding with fH. Table 5.6 illustrates the values of the ratio of K_A/K_S for those windows containing position 138. From the values of K_A/K_S ratio shown in Table 5.5, it was seen that all the windows except for one (positions 133-138), the K_A/K_S ratio was more than 1. This could have been due to some codon starting from position 139 being under positive selection pressure or it could also be that the initial assumption is wrong. Maybe not all fH binding sites are under negative selection pressure. Maybe a few of them are under positive selection pressure.

Table 5.6: Analysis of the K_A/K_S ratio using the sliding window approach for Position 138. Positions where the K_A/K_S ratios are more than 0.9 are highlighted with red colour.

Window of the nucleotide positions	**Corresponding amino acid positions**	K_A	K_A/K_S
397-414	133-138	0.044	0.128
400-417	134-139	0.388	1.131
403-420	135-140	0.434	1.265
406-423	136-141	0.606	1.767
409-426	137-142	0.390	1.137
412-429	138-143	0.472	1.376

Analyzing Amino Acid Position 226

This was the 18th position that takes part in the binding with fH. Table 5.7 illustrates the values of the K_A/K_S ratio for those windows containing position 226. All the values of the K_A/K_S ratio were less than 1 except for one window (positions 223-228) and this could be due to some codon within this range. Hence, position 226 could still be under negative selection pressure.

Table 5.7: Analysis of the K_A/K_S ratio using the sliding window approach for Position 226. Positions where the K_A/K_S ratios are more than 0.9 are highlighted with red colour.

Window of the nucleotide positions	**Corresponding amino acid positions**	K_A	K_A/K_S
661-678	221-226	0.038	0.111
664-681	222-227	0.192	0.560
667-684	223-228	0.360	1.050
670-687	224-229	0.272	0.793

Continued on next page

Table 5.7 – continued from previous page

Window of the nucleotide positions	Corresponding amino acid positions	K_A	K_A/K_S
673-690	225-230	0.089	0.259
676-693	226-231	0.139	0.405

The determination of the positive selection sites using the K_A/K_S ratio was incorporated as one of the constraints while developing the pipeline of programs for generating the new variants. The positive selection sites, since they are prone to mutate more than other positions, had the option to be generated by separate programs and also with other positions with which they might be correlated or co-occurring. Section 5.7 shows how the positive selection sites of the new variants are generated, both separately and with other correlated or co-occurring positions.

Our next step was to find the tertiary structures of all the 190 fHbp sequences using SWISS-MODEL [13, 33, 47, 67] and to view the structures using the Swiss-PdbViewer [20, 45]. To derive the tertiary structures, SWISS-MODEL used templates from the PDB (Protein Data Bank) [24]. In this case, two templates were used: 2W80C and 2W80H (both of which were chains of the fHbp variant 2W80). However, there were instances where another template was used in addition to either of 2W80C or 2W80H. This template, 1YS5, is always a subset of 2W80C or 2W80H, and only a small portion of the fHbp sequence was modeled by 1YS5. As a result, 2W80C and 2W80H were the templates considered for predicting the tertiary structures of fHbp sequences.

Two sequences were selected from the set of 190 fHbp sequences, B6EAW6 and C0JFN4. As a homologous model structure, SWISS-MODEL used PDB structure 2W80C for B6EAW6 and 2W80H for C0JFN4. B6EAW6 had a sequence identity of 94.191% with 2W80C. For C0JFN4 and 2W80H, the sequence identity was 92.116%. Figures 5.7 and 5.8 illustrate the structures of B6EAW6 and C0JFN4, respectively, built using SWISS-MODEL and Swiss-PdbViewer. These two sequences using different structural templates were selected in order to identify the structural differences between 2W80C and 2W80H. For this purpose, the two predicted structures were superimposed and their RMSD calculated using Swiss-PdbViewer [20, 45]. Figure 5.9 illustrates the two superimposed structures. Some differences in the two structures were visually evident, and their RMSD was 0.20 $\mathring{A}$, the value of which indicated that there were some structural differences between 2W80C and 2W80H. Two of the differences are highlighted in Figure 5.9 with red circles. One difference was in one of the beta sheets of the two structures, where the beta sheet in B6EAW6 was shorter than the corresponding beta sheet of C0JFN4. Another difference (red circle at the bottom of the figure) was in one of their coils. Other minor differences are not highlighted in Figure 5.9 due to space limitations.

Figure 5.7: Predicted tertiary structure of an fHbp sequence B6EAW6 which uses the template 2W80C of PDB. The structure was derived using SWISS-MODEL and was visualized using the Swiss-PdbViewer. The structure shows two beta barrels where the beta sheets (illustrated by the arrows) have been coloured in yellow, the coils in grey, and the alpha helix in red.

The tertiary structures act as a constraint for verifying the validation of the new variants to be generated. From the above findings, it was seen that the existing fHbp sequences used two templates to derive their structures. Hence, the new variants in order to be considered as valid should use these two templates to derive their tertiary structures. However, like the existing fHbp sequences, the variants would have variable percentage identity with the template. This is further discussed in Section 5.9. The predicted tertiary structures of the fHbp sequences is significant also because they were used in the next step to determine the energy values and to set the valid energy boundary for the new variants (Section 5.6). New variants having the energy values less than the boundary energy value would be considered as a valid new variant (Section 5.9).

Figure 5.8: Predicted tertiary structure of an fHbp sequence C0JFN4 which uses the template 2W80H of PDB. The structure was derived using SWISS-MODEL and was visualized using the Swiss-PdbViewer. The structure shows two beta barrels where the beta sheets have been coloured in blue, the coils in lime green, and the alpha helix in red.

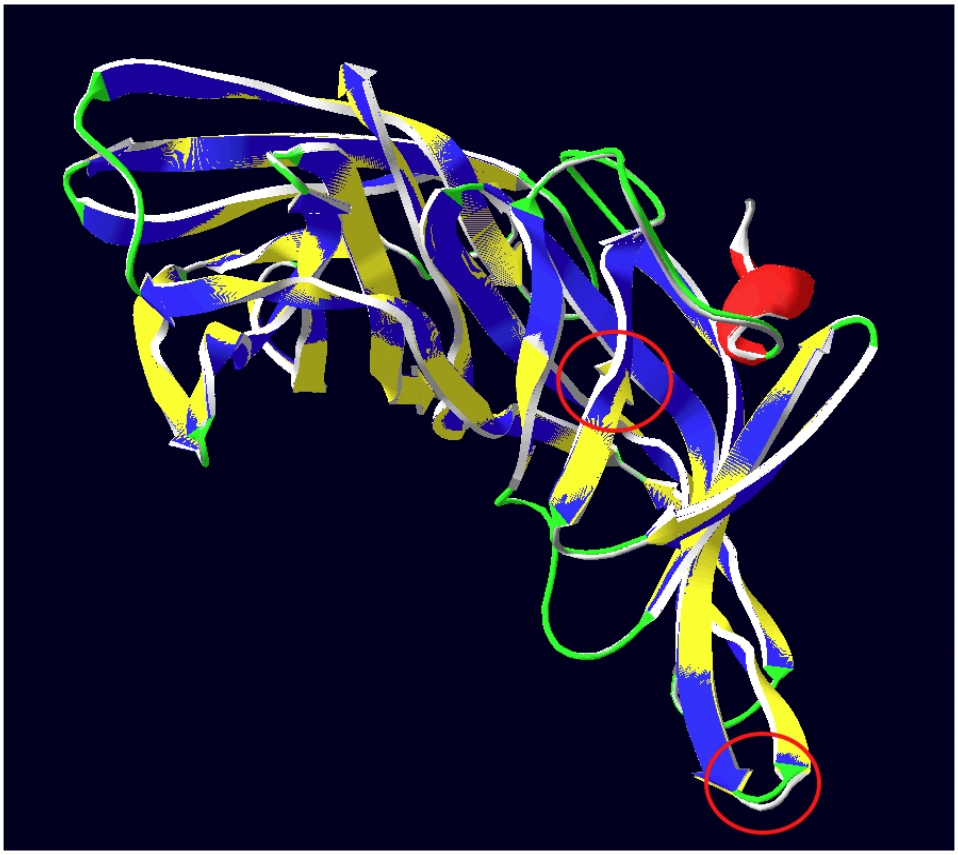


Figure 5.9: Superimposed structures of B6EAW6 and C0JFN4. The structures have been superimposed using Swiss-PdbViewer and there were some structural differences between the two. Their RMSD was 0.20 $\mathring{A}$. The red circles mark two of the structural differences in this figure.

5.6 Determining the valid boundary set by the highest energy of existing fHbp structures

The results for this section were derived as part of Step 9 of our Methodology (Figure 4.1). In this step, we determined the energy values for all the fHbp structures predicted in Step 8. First, the energy of each structure was minimized using Swiss-PDBViewer [20, 45]. This was done because energy minimization can repair distorted geometries as discussed in Section 2.9. Then FoldX [44, 72, 73, 74] was used to determine the energy of each structure. FoldX uses the full atomic description of a protein structure.

Table B.7 in Appendix B give the results of these energy-related operations for all of the 190 fHbp sequences. The first column names the sequence. The second column identifies the template used by SWISS-MODEL. The third column lists the sequence identity of the sequence in the first column with the template listed in the second column. The fourth column provides the energy value initially determined by SWISS-MODEL when the structure was derived. The fifth column lists the energy after energy minimization was performed. Both these energies are in KJ/mol. The sixth and seventh columns are the energy values derived using FoldX before and after energy minimization was performed, respectively. The energy values in these columns are in Kcal/mol.

Table B.7 reveals that the energy values determined by FoldX were considerably reduced by energy minimization for all 190 fHbp sequences. For this research the most important column in the table is the seventh, the FoldX energy after energy minimization. It appears that these energy values tend to increase inversely to the percentage identity expressed in the column 2. In general, the higher the percentage identity, the lower the energy values in column 7. However, this trend is consistent only when comparing sequences whose sequence identities (in column 2) are considerably different. Figure 5.10 shows a line graph which illustrates this trend. The blue line represents the percentage identity and the red line the energy values determined by FoldX after minimization. From the graph it can be seen that as the percentage identity decreases for the existing 190 fHbp sequences their corresponding energy values tend to increase; hence establishing an inversely proportional relationship between the two parameters.

For convenience, some sequences are selected—including the one with the highest energy value in column 7 from Table B.7 in Appendix B—and are given in Table 5.8. The remaining sequences in Table 5.8 were selected to illustrate the variations in the energy values both before and after energy minimization compared to the percentage identity. The following are some observations regarding the information in Table 5.8:

- The percentage identities in rows 1 and 2 are similar. Thus, even though the percentage identity in row 1 is greater than that in row 2, the FoldX energy after minimization (column

Figure 5.10: Line graph illustrating the inversely proportional relationship between the percentage identity and the energy values determined by FoldX after energy minimization is performed on the existing 190 fHbp sequences.

7) is lower in the second row.

- In rows 4 and 5, both the sequences (Q6VRY1 and C0JFK6) have the same percentage identity. Both these sequences had the same energy values after energy minimizations (117.79 Kcal/mol). However, their energy values before minimization were different. In contrast, the sequences in rows 6 and 7 (Q6QCC2 and Q9JXV4) have the same percentage identity, yet the FoldX energies after minimization are different.

- Despite the fact that the sequence in row 8 had lower sequence identity compared to the sequences in rows 6 and 7, this sequence had the lowest energy among the three. However, the differences in the percentage identity between the three sequences were very small.

- The information in rows 8, 9, 10, and 11 (for sequences C0JF89, B2CQ05, C0JF81, and C0JF61) illustrates the trend of increasing FoldX energy values after minimization as percentage identity decreases. In this case the percentage identity for the four sequences varies considerably.

- The sequences in the last two rows of Table 5.8 are another instance where the sequences have similar percentage identity and the sequence with the higher percentage identity had the higher FoldX energy values.

Table 5.8: Energy values determined from the tertiary structures of the existing 190 fHbp sequences.

Sequence	**Template-SM**	**%identity**	**Energy-SM**	**Energy minimization-SPDBV**	**Energy-FoldX**	
					Before minimization	**After minimization**
Q6VRZ1	2W80C	97.095	-11032.109	-11512.100	127.31	114.22
C0JFF9	2W80C	96.680	-10814.681	-11293.024	125.75	112.41
C0JFH9	2W80C	97.925	-10699.144	-11341.228	126.80	110.79
Q6VRY1	2W80C	96.680	-10938.603	-11417.902	131.37	117.79
C0JFK6	2W80C	96.680	-11087.590	-11562.205	130.53	117.79
Q6QCC2	2W80C	100.000	-11121.699	-11593.374	123.58	110.86
Q9JXV4	2W80C	100.000	-11121.699	-11621.195	125.16	117.99
C0JF89	2W80C	99.585	-11200.402	-11669.392	123.66	110.02
B2CQ05	2W80C	95.021	-10834.814	-11485.610	131.07	114.60
C0JF81	2W80C	83.402	-10499.725	-11052.418	145.54	129.67
C0JF61	2W80C	67.635	-10686.848	-11342.188	178.79	143.38
C0JFM5	2W80H	59.016	-7786.667	-9509.900	259.92	212.76
Q19KF7	2W80H	58.607	-9007.075	-9956.911	222.02	189.29

The row highlighted with yellow colour in Table 5.8 identify the sequence with the highest FoldX energy value (after energy minimization) among the 190 fHbp sequences. (The same highlight has been made for Table B.7 in Appendix B.) Sequence C0JFM5 had the highest energy value (212.76 Kcal/mol). Hence, the valid highest energy boundary for structures of fHbp sequences is 212.76 Kcal/mol. This energy boundary was required to validate the new variants after Step 14.

5.7 Generating the new variants

The results of this section were the set of new variants generated using the pipeline of programs developed in Step 10 of the Methodology (Figure 4.1). The individual modules in the pipeline were implemented in Perl.

The programs in the pipeline implemented the amino acid substitution domains for each position, the correlated and the co-occurring relationships and the constraints discussed in Sections 5.3 to 5.5. In total there were 164 Perl programs co-ordinated in the pipeline using UNIX shell scripts. The first program of the pipeline, which required as a command line argument the number of sequences to be generated, outputted the amino-terminal repetitive region, and the invariant segments that flank the five modular variable regions (V_A to V_E). It also produced the template for all the variant sequences. The remaining programs were designed and implemented so that they could be executed in any arbitrary order; in particular, their order of execution could be changed. However, exploring the effects of changing the order of execution was beyond the scope of this thesis.

There was one program in the pipeline written to denote the negative selection sites as 'n' and the positive selection sites as 'p' (Algorithm 4.2). Another separate program was written to mark the 'gap' symbols, where required, along the sequence alignment.

There were 21 separate programs written to realize the positive selection sites of the new variants. This was done because positive selection sites were the regions which were more likely to change compared to other regions. The positive selection sites are all highlighted with red colour in Table B.6 in Appendix B. Next it was seen that there were correlated or co-occurring relationships among some of the positive selection sites. Among the 21 programs, 6 incorporated such constraints while realizing the positive selection sites. The remaining 15 programs were used for positive selection sites which had no such relationships.

The next task was to instantiate the remaining positions, which were either negative selection sites or were neither positive nor negative selection sites and 140 separate programs were written to realize these positions. There were six instances of co-occurrence and one instance of correlation as mentioned in Section 5.4. These individual instances were incorporated in separate programs; that is, one program for one instance and so on. Hence, 6 separate programs were written to realize

the 6 instances of co-occurrence (both "partially" and "fully" co-occurring) and one program was written to realize the single instance of correlation in the sequence alignment. Positions which had no such relationships were instantiated using individual separate programs; that is, one program for generating amino acids for a single position with no correlation or co-occurrence. There were 133 separate programs written to realize such positions. However, it was noticed that there were some positive selection sites (already instantiated by the 21 programs mentioned previously) which had correlated or co-occurring relationships with some of the positions instantiated in this phase. For such cases, the positive selection sites which had correlated or co-occurring relationships with these positions were again included in the programs of this later phase. However, based on the following criterion (already undertaken in Step 10 of Section 4.2),

if the amino acid is already set for a particular position, then we keep that setting

all the algorithms (except for the algorithm of the first program in the pipeline) based on which the programs of this pipeline were developed, checks whether a position to be marked was either 'X' or 'n' or 'p' (see Algorithms 4.1 to 4.6) and remains unchanged otherwise. This meant that the positive selection sites once instantiated by the first 21 programs would not be replaced by the programs of the later phase, even if some of them had correlated or co-occurring relationships with the positions instantiated by the later phase. Hence, in order to accommodate both the phenomenon, where in one case the positive selection sites were generated separately and in the other with the positions which were not positive selection sites, two sets of variants were generated in this research. For the first set of new variants, 200 were generated, where the positive selection sites were instantiated by the programs of the later phase. For this phenomena, the 21 programs which instantiated the positive selection sites were placed after the 140 programs of the later phase in the pipeline. For the second set of new variants, 100 were generated, where the positive selection sites were instantiated by the 21 programs of the first phase; that is, these programs were now placed before the 140 programs in the pipeline. For convenience and better understanding of this thesis, the two sets are termed Set A (with 200 new variants) and Set B (with 100 new variants).

Figures C.7 to C.12 show portions of the new variants of Set A. The figures illustrate the characteristics of the new variants. The amino-terminal repetitive regions, the invariant segments and the five modular variable regions are highlighted with the same colour as was used for the 190 fHbp sequences (illustrated in Figures C.1 to C.6). Note that the new variants lack the signal protein part and start from the mature protein part; that is, the amino-terminal repetitive segment. The consensus sequence illustrated in the figures was generated by the EMBOSS *prettyplot* program from the set of 200 new variants.

5.8 Studying the new variants and filtering of duplicates

This section discusses the results generated in Step 11 of the Methodology of this research (Figure 4.1). After the generation of the two sets of new variants, Set A and Set B, the task was to check whether the pipeline of programs generated any duplicate sequences. In order to verify whether both the sets had unique variants, CD-HIT was used at 100% sequence identity on both the sets. The results obtained showed that there were no duplicates in either set.

The next task was to determine whether any of the new variants were duplicates of the 190 fHbp sequences. To do this, the 190 fHbp sequences were merged first with Set A, and then in a separate test, with Set B. CD-HIT at 100% sequence identity was applied to the resultant merged sets. In both cases it was determined that none of the new variants were duplicates of the original 190 sequences. Hence, it was confirmed that both sets had unique new variants.

Next, it was determined whether the amino acid substitution domains for the variants in Sets A and B were the same as for the original 190 fHbp sequence set. Scripts A.2 and A.3 in Appendix A were applied, and the results generated are included in Tables B.2 and B.3 in Appendix B. The results in Table B.2 show that the frequency of the different amino acids in each position along the alignment of the new variants in both sets was in accordance with the frequencies from the original 190 sequences. The results in Table B.3 confirm that gaps are generated in the correct positions in both sets of new variants.

Finally, it was desirable to examine the new variants in Sets A and B to determine if they had the same positive and negative selection sites as the original 190 fHbp sequences. This was performed using the K_A/K_S ratio as described in Section 5.5. A hurdle in this procedure was that calculating $K_A$and K_S required the nucleotide sequences for the protein sequences. Since the sequences in Sets A and B were synthetic, such nucleotide sequences were not available. Synthetic nucleotide sequences could have been generated using tools such as the EMBOSS *backtranseq* [3] program. However, certain issues would have arose. For instance, due to the ambiguity in the genetic code, it could not be known for certain what nucleotide sequence would correspond to a particular "new variant". Even if *backtranseq* were provided with a table of codon usage frequency for *N. meningitidis*, it could still not be known for certain what nucleotide sequence would correspond to a particular "new variant". If *backtranseq* chose a nucleotide which was different, then this in turn would have affected our K_A/K_S value and the detection of positive and negative selection sites. Hence, it was decided not to proceed with this part of verification of the new variants.

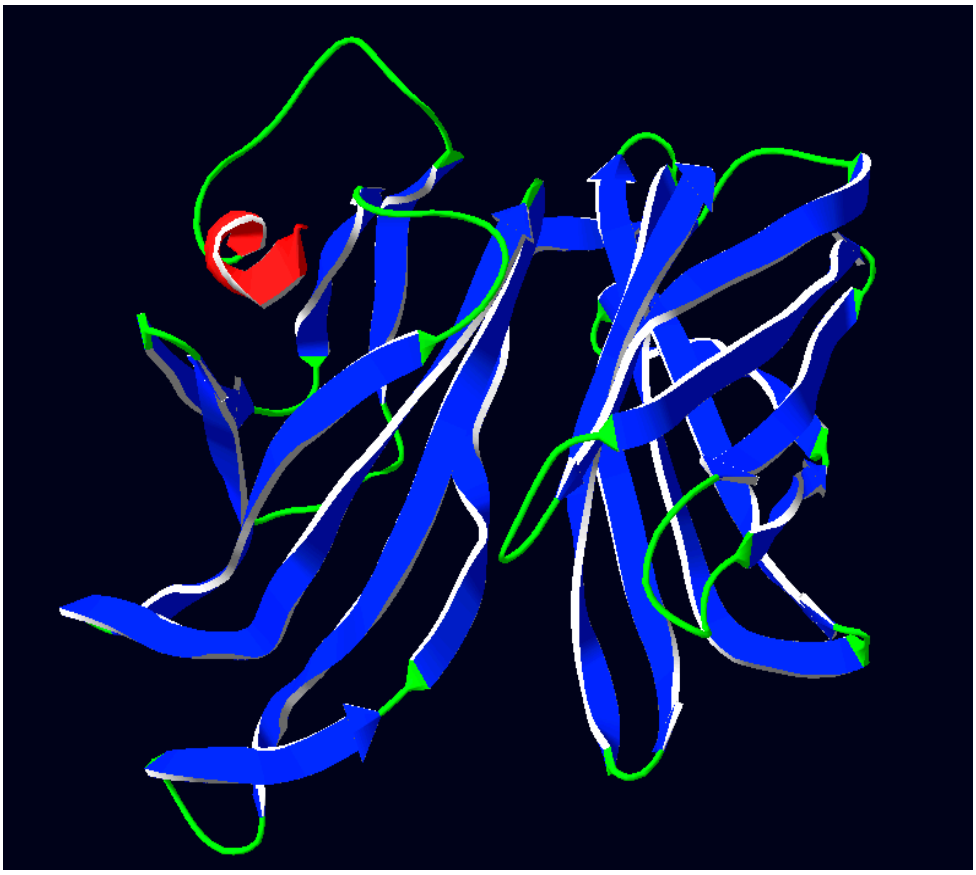


Figure 5.11: Predicted tertiary structure of a new variant of Set A which uses the template 2W80H of PDB. The structure was derived using SWISS-MODEL and was visualized using the Swiss-PdbViewer. The structure shows two beta barrels where the beta sheets have been coloured in blue, the coils in lime green, and the alpha helix in red. The same structural patterns were observed for the existing fHbp sequences (Figures 5.7 and 5.8).

5.9 Determining the validity of the new variants

This section discusses the results generated in the loop which started from Step 12 and involved Steps 13 to 15 (Figure 4.1). The purpose of the loop was to determine the validity of the new variants based on their predicted tertiary structures and the energy values of those structures.

The first task in the loop was to determine the structures of the new variants (Step 13, Figure 4.1). For all of the new variants of both the sets, their tertiary structures were found using SWISS-MODEL [13, 33, 47, 67] as was done for the original 190 fHbp sequences in Step 8. The structures were visualized using Swiss-PdbViewer [20, 45]. It was observed that, similar to the original sequences, the new variants used three templates from PDB: 2W80C, 2W80H, and 1YS5. For the same reasons explained in Section 5.5, the structures using the templates 2W80C and 2W80H were taken into account, and the ones based on 1YS5 were discarded. Figure 5.11 shows a tertiary structure of one of the new variants of Set A, which uses the template 2W80H of PDB.

The next task in the loop was to determine the energy values of these tertiary structures for all the variants in Sets A and B. The same procedure was followed as for the 190 fHbp sequences. First, energy minimization was performed using Swiss-PdbViewer. Then FoldX [44, 72, 73, 74] was used to determine the energy. Tables B.8 and B.9 give the results. The information in these tables is organized in the same manner as in Table B.7, with the first column depicting the name of the sequence; the second column, the template used in SWISS-MODEL; the third column, the percentage identity of the sequence in the first column with the template used in the second column; the fourth column, energy values determined by SWISS-MODEL (in KJ/mol); the fifth column, energy minimization values generated as a result of energy minimization performed on the sequences by the Swiss-PdbViewer (KJ/mol); the sixth and seventh columns, energy values (Kcal/mol) derived using FoldX before and after energy minimizations of the structures, respectively.

A notable reduction in the energy values determined by FoldX is evident in both Tables B.8 and B.9 after the energy minimization on the structures. The same was observed for the original set of fHbp sequences (Section 5.6). As in Step 14 (Section 4.2), the methodology focused on the energy values determined after energy minimization (column 7). The highest value in column 7 has been highlighted with yellow colour in both the Tables B.8 and B.9. It was observed in both tables that these energy values tend to increase inversely to the percentage identity expressed in column 2; that is, the higher the percentage identity, the lower the energy values in column 7. However, this trend is consistent only when comparing sequences whose sequence identities (in column 2) are considerably different. This observation is illustrated in Figures 5.12 and 5.13 for Sets A and B respectively. The same was seen in Table B.7 and Figure 5.10 for the original set of sequences (see Section 5.6 for explanation).

In Table B.8, which lists the energy values for Set A (200 new variants), the highest energy was 195.52 Kcal/mol (variant 52). For Table B.9, which lists the energy values for Set B (100 new variants) the highest was 202.62 Kcal/mol (variant 39). In Step 9, it was determined that the valid boundary set by the highest energy value for fHbp sequences was 212.76 Kcal/mol (Section 5.6). All the energy values for all the new variants in both the sets were less than 212.76 Kcal/mol. Hence, all the new variants in both sets were judged as valid new variants of fHbp sequences. The Sets A and B contained new variants that nature might "allow" but which have not as yet appeared.

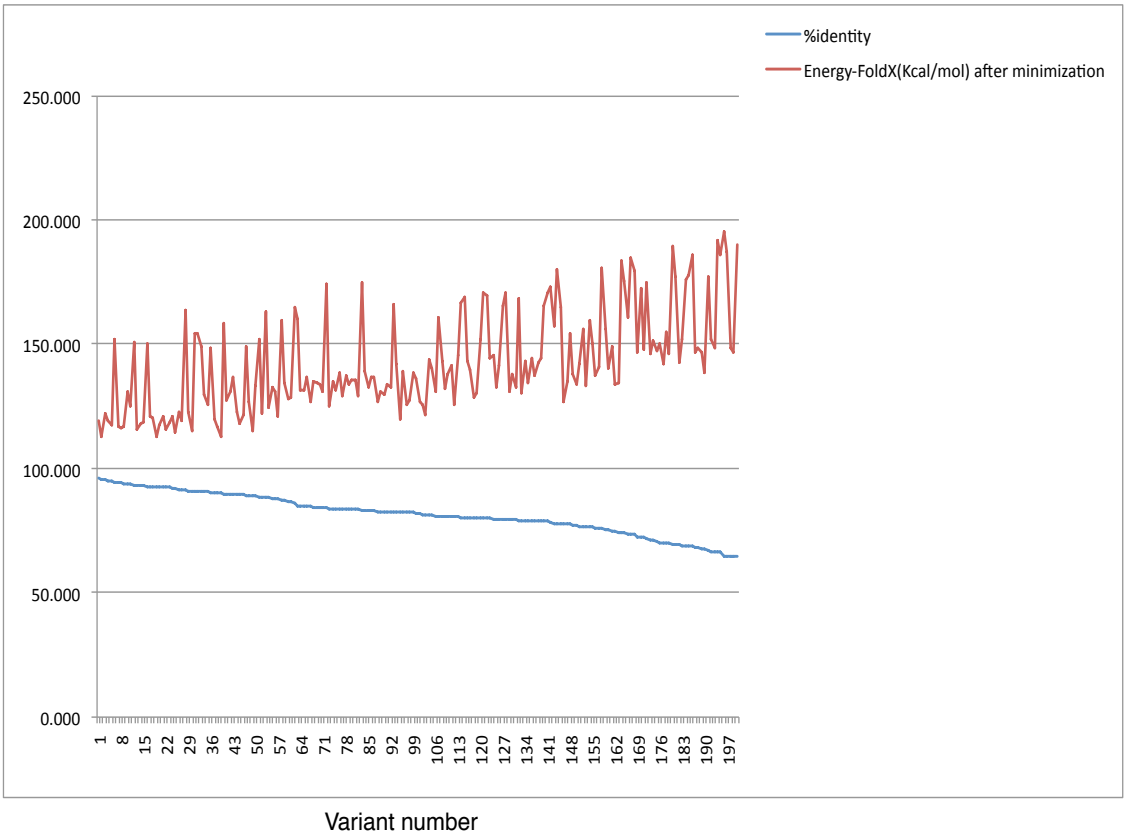


Figure 5.12: Line graph illustrating the inversely proportional relationship between the percentage identity and the energy values determined by FoldX after energy minimization is performed on the 200 variants of Set A.

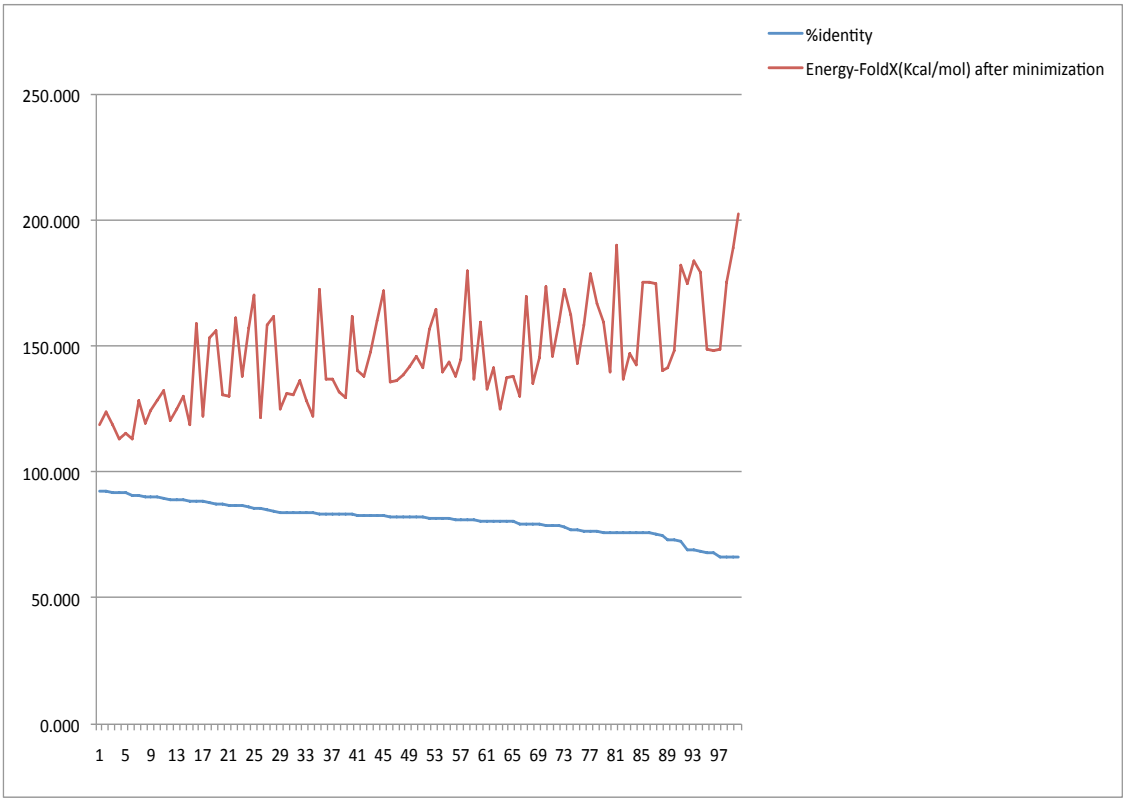


Figure 5.13: Line graph illustrating the inversely proportional relationship between the percentage identity and the energy values determined by FoldX after energy minimization is performed on the 100 variants of Set B.

Chapter 6

Discussion

This research aimed to predict the evolution of fHbp, an antigen present on the surface of all strains of *Neisseria meningitidis*. For this purpose, unique sequences of fHbp were studied in order to determine their various characteristics including the correlated or co-occurring regions in the sequence alignment, and the energies of their predicted structures. Based on all these characteristics new variants of fHbp were generated using the pipeline of programs developed as part of this research. These new sequences were the variants of fHbp that nature might "allow" but which have not as yet appeared.

This chapter summarizes the results presented in Chapter 5. It discusses the conclusions drawn from the experiments performed in our research and also highlights some of the future work that could be done later to further enhance the results.

6.1 Conclusions & Remarks

As mentioned in Chapter 3, the goal of this research was to generate valid new variants of fHbp sequences and there were certain objectives required to be achieved in order to meet this goal.

- The first objective, which was a significant one, led to the development of an extensible methodology for this research, which would form the foundation for more work in future.

- The second objective led to the study and determination of the characteristics of the existing fHbp sequences. The existing fHbp sequences were downloaded from UniProtKB and their characteristics were studied. It was observed that the sequences had specific variant regions separated by invariant segments. In addition, the domains of possible amino acid substitutions for each variable position showed that variation in these regions was restricted.

- The determined constraints showed that there were correlation and co-occurring relationships amongst various positions in the variable regions of the sequences. In addition, there were positive selection sites which were prone to change and negative selection sites which tended to be conserved.

- SWISS-MODEL was used to predict the tertiary structures of the fHbp sequences and Swiss-PdbViewer was used to visualize them. It was seen that SWISS-MODEL used three templates from PDB to predict the structures: 2W80C, 2W80H, and 1YS5, of which only a small portion of the fHbp sequence was modeled by 1YS5 . The structures had some structural differences between them, as discussed in Section 5.5.

- As part of achieving the next objective, the energy of the tertiary structures were found using a widely-used software (FoldX). FoldX was used because it uses a full atomic description of the structures of proteins.

- The next objective—a significant one—was to develop a pipeline of programs for generating the new variants of fHbp. For this, the Perl programming language was used to write the programs and UNIX shell script for coordinating the programs in a pipeline. As part of the design goal, the pipeline of programs were built such that there was flexibility in the order of execution of individual programs; that is, the order of the Perl scripts within the pipeline could be varied except for the first program in the pipeline. Another design goal achieved was that additional Perl scripts could be added to the pipeline; that is, there was no limit to the number of programs that could be added, hence, forming a platform for future work.

- The last objective was to determine the validity of the new variants generated by the pipeline of programs. For this the characteristics of the new variants were studied and compared with those of the existing fHbp sequences. This was done in order to determine whether the variants possessed the same characteristics as those of the existing ones. It was also determined whether the same constraints and correlation or co-occurring relationships existed in the new variants. The tertiary structures of the new variants were predicted and it was found that SWISS-MODEL used the same templates as were used for the existing ones. Their energy values were determined using FoldX and then compared with those of the existing fHbp sequences.

A more reliable check of the viability of the variants could have been carried out in a wet lab. However, it was out of the scope of this research; hence it can be carried on as part of future work for this research.

The results in Chapter 5, specifically Section 5.9, indicated that all the new variants generated had possible structures with energy values less than the highest energy value (212.76 Kcal/mol), which was set as the valid boundary of energy for existing fHbp sequences (Section 5.6). This meant that the sets of new variants were all judged as valid, which was an unexpected result. At least some variants whose energy values were not within the valid range were expected due to the stochastic nature of the generation process. The following are some possible reasons for this result.

i. There were issues regarding the analysis of the fHbp binding sites to see if they were all under negative selection pressure. Some of the binding sites had conflicting K_A/K_S ratio value indicating that all the binding sites might not be under negative selection pressure. This might be indicative of a problem with the determination of sites under negative selection pressure, or parts of the methodology.

ii. Different positions of the aligned 190 fHbp sequences were visually examined to determine if the positions of the restricted variable positions had correlated or co-occurring relationships. Since co-occurrence and correlation constraints were determined visually, and the situation being analyzed was complex (263 columns and 190 rows), it is quite conceivable that certain key constraints of this type were over-looked.

iii. There were types of constraints that were identified but not exploited; the constraints were deemed to be outside the scope of this research. Examples of such constraints were situations where one amino acid has higher possibility of substituting another amino acid or occurrences of intermediate residues in possible mutations (see the two examples illustrated in the *Molecular Evolution* section of Step 8 Section 4.2). Such constraints may have determined that some of the new variants would not be allowed by nature or could have made the similarity boundary—both in structure and energy levels—even tighter.

iv. The 3D molecular structures predicted by SWISS-MODEL may have been erroneous or unrealistic due to a lack of model structures available at PDB. This could have led to misleading or erroneous energy values for the existing sequences or new variants. Even though PDB had a number of fHbp structures available – e.g. 2W80 (composed of 8 chains), 2W81 (6 chains), 2KDY (1 chain), 2KC0 (1 chain), 1YS5 (1 chain) – Tables B.7 through B.9 indicate that only two two models (chains C and H of 2W80) were used.

v. The number of variants generated may have been too small. There might have been some variants with energy values which did not fall within the valid energy range if a larger set of new variants were generated; for instance, a thousand new variants.

vi. Despite the testing of each program in the pipeline and studying individual results generated in each phase, there might have been errors in the pipeline of programs developed for generating the new variants, made either when gathering characteristics from the existing sequences or when implementing programs to manifest those characteristics.

As a conclusion, it can be said that if the variants satisfies all the constraints mentioned in Chapter 4.2 and if they are generated by the pipeline of programs, then this thesis provides evidence they will likely fit into the allowable energy values. What cannot be concluded is the converse: that if a variant does not satisfy the constraints, that it will not be allowable. Developing a more precise

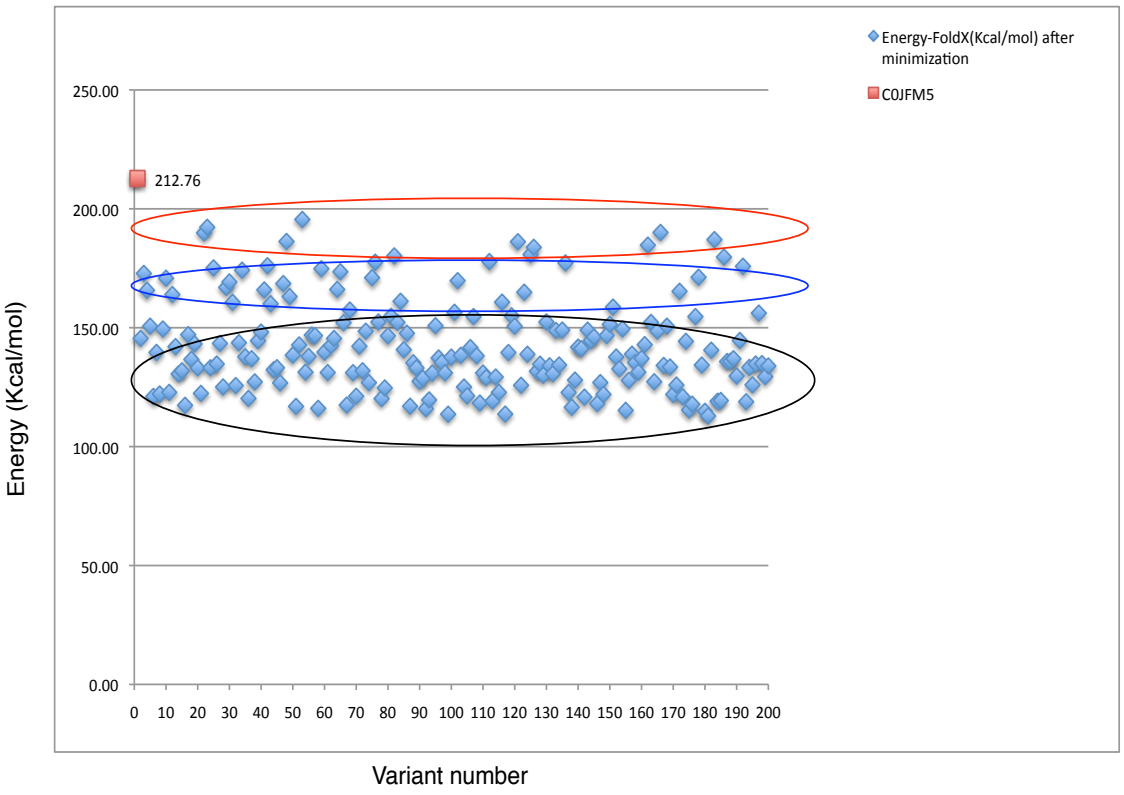


Figure 6.1: Scatter plots illustrating energy (Kcal/mol) of the 200 new variants of Set A generated by FoldX after energy minimization is performed. Energy of C0JFM5 (212.76 Kcal/mol) is the highest energy recorded of the existing fHbp sequences, hence forming the boundary for valid energy values for the new variants.

characterization is therefore left for future work. Finally, it could be that, in fact, all of the 300 new variants would be allowed by nature (were "valid") and that the energies are simply reflecting this fact. The only way to make a definitive determination in this respect is through "wet lab" verification.

However, it is possible to make a probabilistic conclusion regarding the probability of certain variants occurring more than other variants from the distribution of the energy values. For this the energy values of both sets A and B are plotted in the energy distribution graphs in Figures 6.1 and 6.2, respectively. From Section 5.6 it has been shown that the higher the percentage identity, the lower the energy values; hence, variants with lower energy values having higher percentage identity will be more likely to occur than those with higher energy values and lower percentage identity.

In Figure 6.1 the scatter plots illustrating the energy of the 200 new variants of Set A are highlighted by the blue dots and the energy value of C0JFM5 (212.76 Kcal/mol) which is the highest energy recorded of the existing fHbp sequences is marked by the red square. This red square forms the boundary for the maximum energy that the new variants may have. Since all

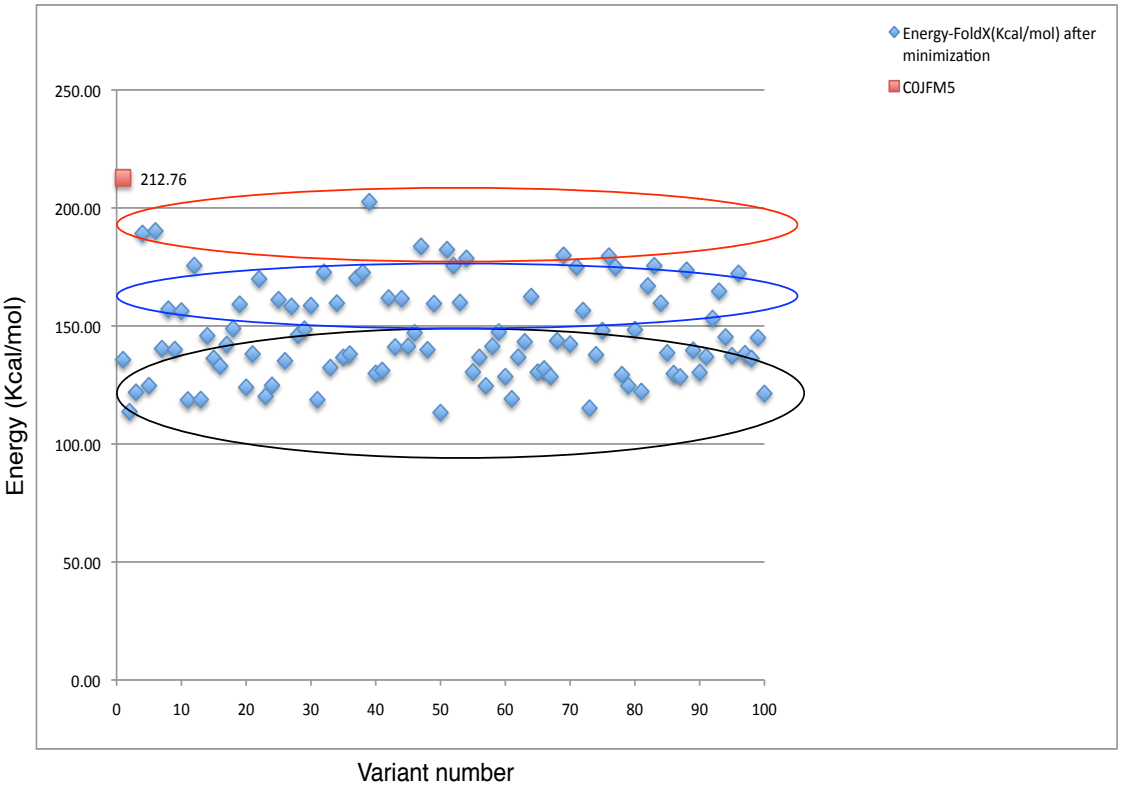


Figure 6.2: Scatter plots illustrating energy (Kcal/mol) of the 100 new variants of Set B generated by FoldX after energy minimization is performed. Energy of COJFM5 (212.76 Kcal/mol) is the highest energy recorded of the existing fHbp sequences, hence forming the boundary for valid energy values for the new variants.

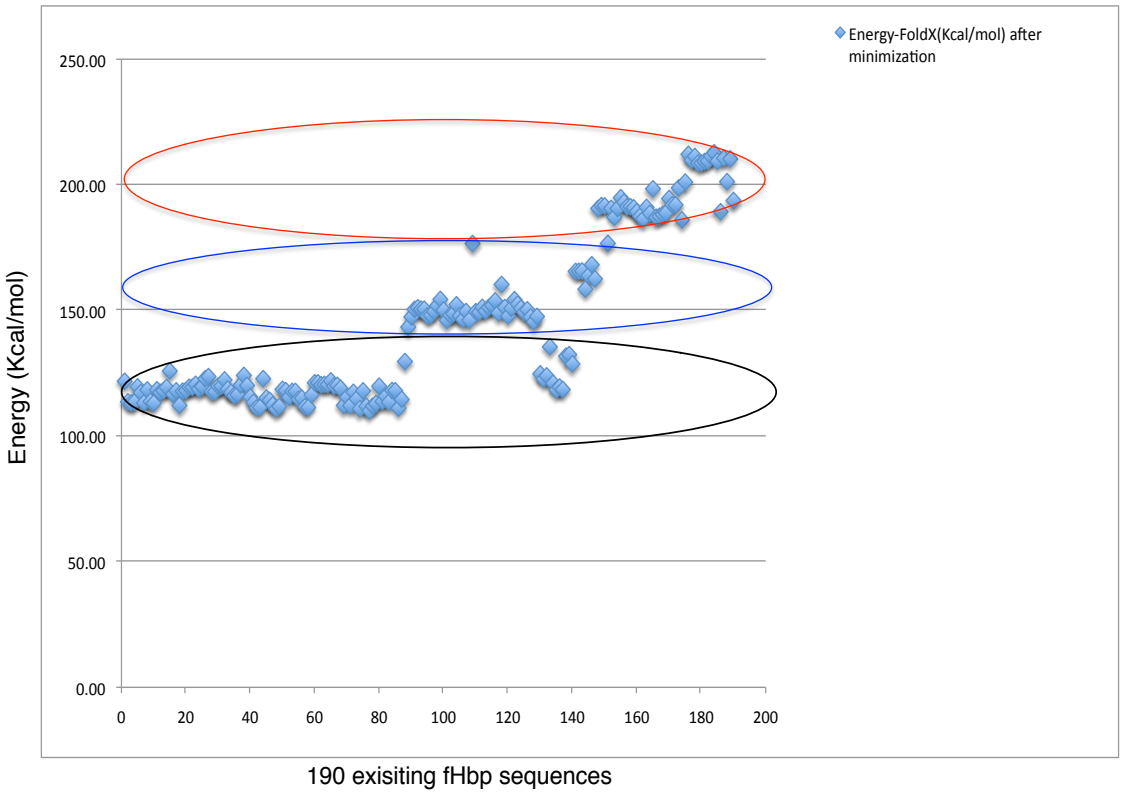


Figure 6.3: Scatter plots illustrating energy (Kcal/mol) of the existing 190 fHbp sequences generated by FoldX after energy minimization is performed.

the energy values were less than this highest value, it is possible to make a probabilistic conclusion from the energy distribution graph and mark those variants which were more likely to occur and those that were less likely to occur. The first region marked by the black oval shape encircles the energy values of those variants which were far away from the boundary energy and were more likely to occur. It can be seen that this encircled region is more dense when compared with the other two encircled regions. The second region marked by the blue oval shape encircles those variants which have moderate chance of occurrence. The last one encircled by the red oval shape and closer to the boundary energy marks those variants which are less likely to occur. This region is the least dense covering approximately around 11 new variants. This constitutes around 5.5% of the 200 new variants in Set A. The last column in Table B.8 illustrates which variant belongs to which of the three regions.

The same probabilistic study was carried out on the energy values of the 100 new variants of Set B, illustrated by Figure 6.2. The same blue dots mark the energy values of the new variants and the red square the energy value of C0JFM5 forming the boundary for the maximum energy the 100 new variants may have. The first region marked by the black oval shape are the energy value of those variants which are far away from the boundary energy plus it is the most dense region. The blue oval shape marks those variants which have moderate chance of occurrence. The last shape illustrated by the red oval shape encircles the energy values of those variants which are closer to the boundary energy value and are less likely to occur. This region covers 8 new variants approximately, constituting around 8% of the 100 new variants in Set B. The last column in Table B.9 illustrates which variant belongs to which of the three regions.

Figure 6.3 illustrates the scatter plot of the energy of the 190 existing fHbp sequences. In this graph the same kind of energy distribution is observed with the distribution being divided into three regions. However, the region encircled by the red oval shape has more numbers of fHbp sequences than were observed for both plots in Figures 6.1 and 6.2.

Based on the scatter plots of Figures 6.1, 6.2, and 6.3 histograms shown in Figures 6.4, 6.5, and 6.6 were built respectively. The histograms illustrates the energy distributions of the two sets of variants (Figures 6.4 and 6.5) and the existing 190 fHbp sequences (Figure 6.6). For all the histograms the energy range has been divided onto a scale of 20 Kcal/mol. It was observed that for both the sets of variants, the maximum number of variants had energy values ranging between 120-140 Kcal/mol, whereas for the existing fHbp sequences the maximum number of sequences had energy values ranging between 100-120 Kcal/mol. There were no variants in Set A which had energy value of more than 200 Kcal/mol, hence that bar is absent in Figure 6.4. For Set B there were minimum number of variants which had energy values ranging between 200-220 Kcal/mol (Figure 6.5). However, for the existing fHbp sequences, moderate number of sequences had energy values within 200-220 Kcal/mol range (Figure 6.6). Based on the energy distributions depicted by

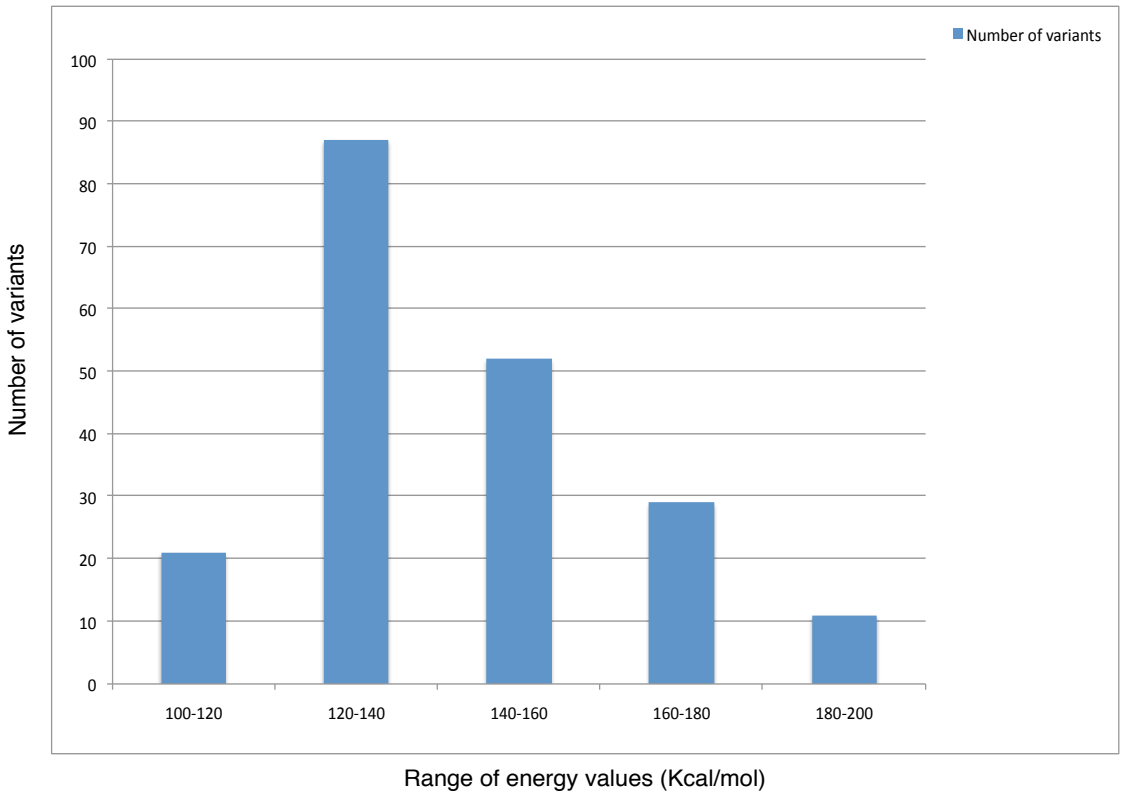


Figure 6.4: Histogram illustrating energy distribution of the 200 new variants of Set A.

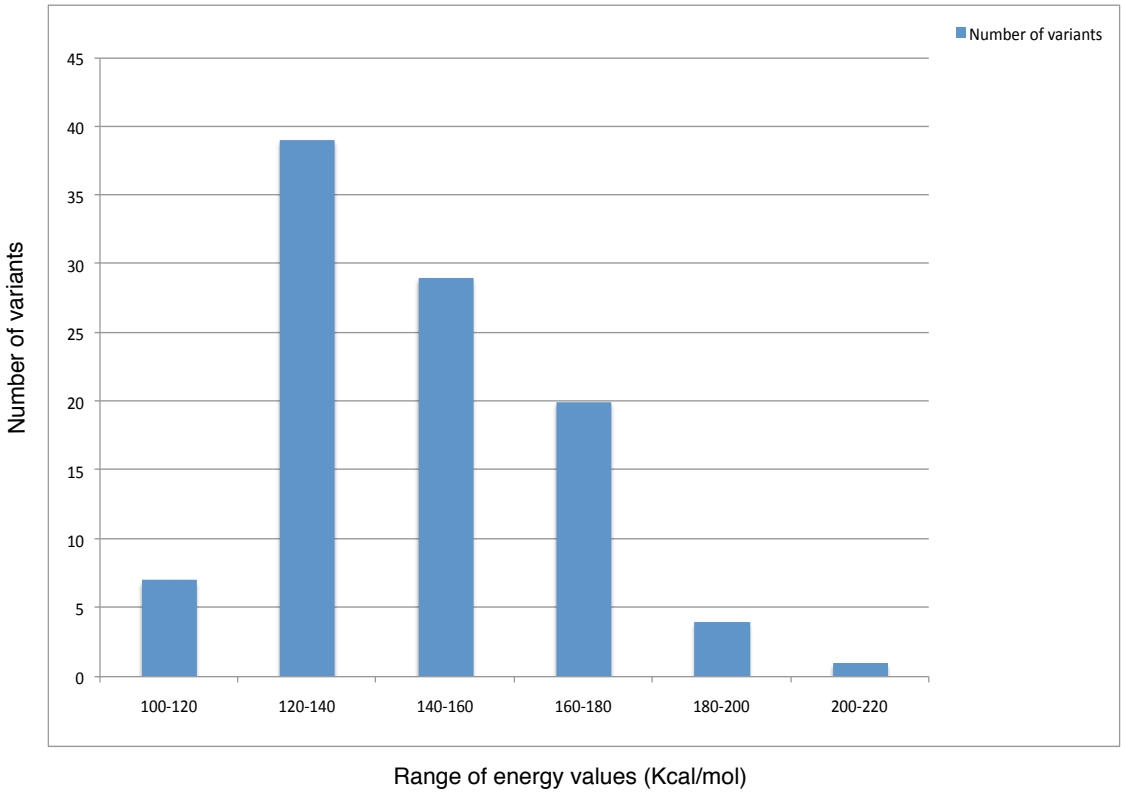


Figure 6.5: Histogram illustrating energy distribution of the 100 new variants of Set B.

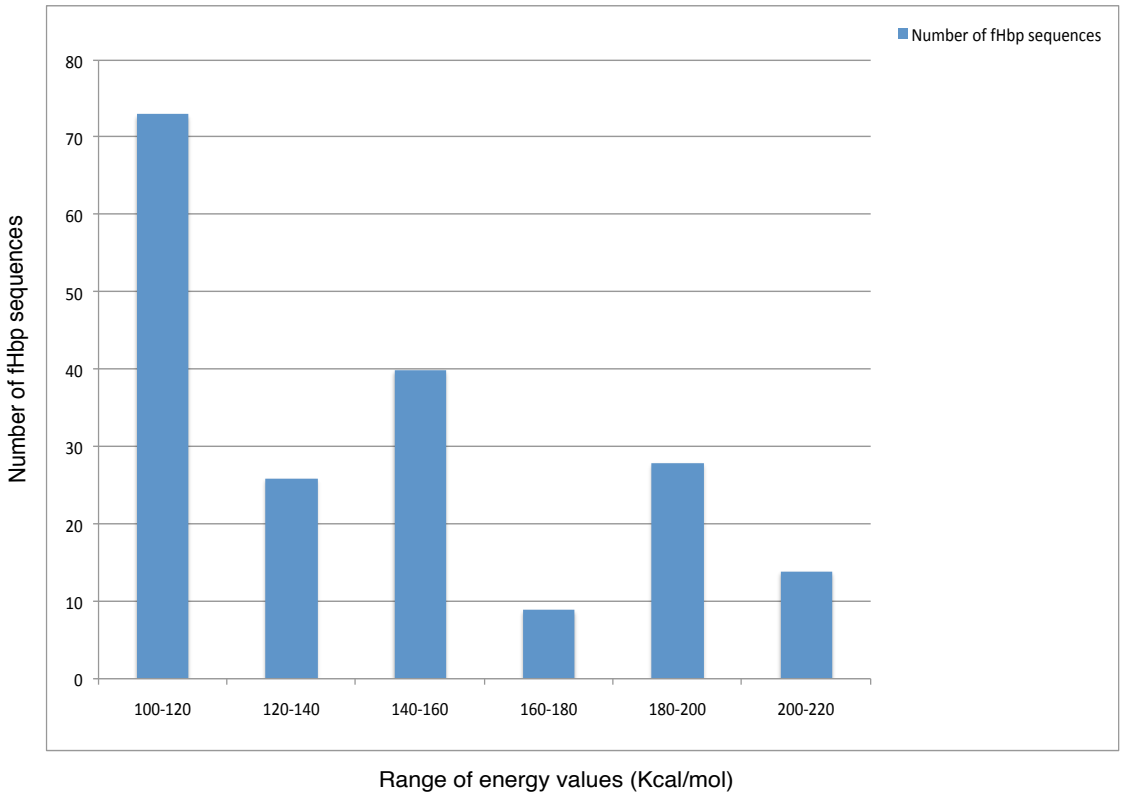


Figure 6.6: Histogram illustrating energy distribution of the existing 190 fHbp sequences.

the histograms of Figures 6.4, 6.5, and 6.6, it can be concluded that the number of variants as well as the existing fHbp sequences were more for ranges with the lower energy values than for ranges with the higher ones.

The last discussion on this thesis is based on to what extent this research has tried to capture realistic variants of *Neisseria meningitidis*, that is, how confident one can be that the variants obtained through this research were functionally meaningful. There were some constraints that were functionally meaningful and some which were superficial in nature. However, the approaches undertaken in this research is just an early step towards generating realistic variants and requires much more work in future. For instance, this research due to its limitations was not able to investigate affinity between the 21 binding sites of fH and fHbp. As suggested by committee member, Dr. Nate Osgood, one critical "fitness function" that could be worked on in future would be the affinity of the 21 binding sites of fHbp to fH. A simulation approach can be undertaken involving incremental generation of variants, that is, one mutation at a time and discarding at every step those that don't have high affinity to fH. This simulation process can be repeated several times until the variants that might seem sufficiently fit emerges. Such simulation process will result in variants that has passed the affinity test and will be more functionally meaningful in nature.

The literature was probed to find other similar research works involving the prediction of the evolution of the factor H binding protein antigen. Unfortunately, similar works could not be found. However, some works were found which involved study of the fHbp antigen and vaccine development. Most of these works were highly concentrated on biochemistry and not on bioinformatics. Only one work involved some bioinformatics-based research in addition to biochemistry. The title of this paper is *"Sequence Diversity of the Factor H Binding Protein Vaccine Candidate in Epidemiologically Relevant Strains of Serogroup B Neisseria meningitidis"* [59]. This research focused only on the strains of serogroup B. The sequences collected were divided into two subfamilies, A and B, based on the sequence diversity, unlike our work, where no such classifications were made. The method used in this research involved Multilocus sequence typing (MLST) analysis. The bioinformatics portion of this research involved aligning the sequences using ClustalW, just as was done here. However, in addition with aligning, distances were also calculated using the neighbor-joining method in ClustalW, which was not done in our work. Consensus sequence were calculated using EMBOSS program *cons*, similar to what was done for our work. Minimum spanning trees were constructed for both the subfamilies A and B in this research, unlike our work. The fHbp sequences were studied to identify the conserved and the variable regions, similar to our work. However, no regions which had co-ocurring or correlational characteristics were identified in this research work. The concluding remark for this research based on their results was that the diversity of the strains that were observed implied that studying the distribution of the vaccine antigen itself was more

important than relying on some epidemiological substitution like MLST.

6.2 Future Work

The following is future work that could be carried out to improve on or extend this work.

i. Testing, analyzing, and verifying of the new variants by biochemists using wet lab techniques. This is perhaps the most significant future work that can be carried out for further determination of the validity of the new variants.

ii. More sequences have been deposited at UniProtKB, so the research could be repeated, but taking advantage of the updated information at UniProtKB.

iii. Step 7 (Section 4.2), which determines the correlation and co-occurrence relationships, can be performed by a statistical or a computational tool, rather than being performed visually. Since it was performed visually, some correlations and co-occurrences could have been missed.

iv. Implement the complete characteristics of *Molecular Evolution* as discussed in Step 8 (Section 4.2), so that generation of new variants will make use of those amino acids that have higher possibility of substituting another amino acid.

v. Include the chemical properties of the amino acids as another type of constraint when generating the new variants.

vi. While determining the spatial and physicochemical constraints in Step 8 (Section 4.2), the study of the tertiary structures started with the correlated and co-occurring positions in region V_A. In future, the study can start with positions in other regions.

vii. Include both the tertiary structural constraints as mentioned in *Spatial and physicochemical constraints* of Step 8 (Section 4.2):

(a) amino acids that must be within certain distances of each other to form, for example, an active site;

(b) amino acids with conflicting physicochemical properties cannot be within certain proximities of each other.

viii. The possibility of using phylogenetic analysis to provide constraints to drive the generation of variants.

ix. Use programs other than SWISS-MODEL to determine the structures of the original sequences as well as the variants. Of particular interest would be using programs which are not restrained by the lack of availability of model structures at PDB.

x. Use a commercial quality energy minimization program, such as Amber [1, 49] to perform the energy minimization. Amber is available for research use on WestGrid [28].

xi. There were two alternative options for generating positions in the new variants with interference problems (Step 10 of Section 4.2). For such positions, the first alternative was implemented in this research. The second alternative – to allow subsequent steps to modify an already placed amino acid according to the constraints at that stage – can be implemented.

xii. Rearranging the programs in the pipeline should result in different variants being produced. While this might be desirable for the user, it is inefficient since it requires manual modification of the pipeline. To increase the efficiency, a higher level to the pipeline can be developed, where the shuffling of the programs in the pipeline is done using an algorithm instead of shuffling manually by the user as is done now or the user can be invoked to determine the shuffling order. Such facility to programmatically control the execution of the modules in the pipeline may lead to the generation of more diverse variants.

xiii. To look at the energy in just the binding pockets of fHbp, that is, at the 21 binding sites.

xiv. To add a control measure to the research, such that variants would be generated which did not use the constraints or the correlation and co-occurrences properties. The restricted variable regions of these variants would just have randomly mutated amino acids and then it could be verified whether these variants had the same tertiary structures and whether they still had energy values below the valid boundary energy, or whether they had higher energy.

References

[1] http://ambermd.org/. Accessed March 2011.

[2] http://download.cnet.com/KaKs-Calculator/3000-2383_4-75305712.html. Accessed October 2010.

[3] http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/backtranseq.html#input.1. Accessed September 2010.

[4] http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/cons.html. Accessed December 2009.

[5] http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/prettyplot.html. Accessed October 2009. .

[6] http://en.wikipedia.org/wiki/Antibody. Accessed May 2011.

[7] http://en.wikipedia.org/wiki/Genome_project#Genome_annotation. Accessed May 2011.

[8] http://en.wikipedia.org/wiki/Molecular_evolution. Accessed October 2010. .

[9] http://en.wikipedia.org/wiki/Posttranslational_modification. Accessed September 2010.

[10] http://en.wikipedia.org/wiki/Root_mean_square_deviation_(bioinformatics). Accessed December 2010.

[11] http://en.wikipedia.org/wiki/Signal_peptide. Accessed November 2010.

[12] http://jmol.sourceforge.net/. Accessed October 2010.

[13] http://swissmodel.expasy.org/. Accessed January 2010.

[14] http://www.bioinformatics.org/cd-hit/. Accessed June 2009.

[15] http://www.cbs.dtu.dk/services/CPHmodels/. Accessed October 2010.

[16] http://www.cbs.dtu.dk/services/SignalP-3.0/output.php. Accessed December 2009.

[17] http://www.cbs.dtu.dk/services/SignalP/. Accessed December 2009.

[18] http://www.ddbj.nig.ac.jp/FT/full_index.html#mat_peptide. Accessed September 2010.

[19] http://www.ebi.ac.uk/Tools/msa/clustalw2/. Accessed October 2009.

[20] http://www.expasy.org/spdbv/. Accessed June 2010.

[21] http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/. Accessed October 2010.

[22] http://www.megasoftware.net/. Accessed October 2010.

[23] http://www.path.org/. Accessed December 2010.

[24] http://www.pdb.org/pdb/home/home.do. Accessed June 2009.

[25] http://www.pymol.org/. Accessed October 2010.

[26] http://www.umass.edu/microbio/rasmol/. Accessed October 2010.

[27] http://www.uniprot.org/help/uniprotkb. Accessed June 2009.

[28] http://www.westgrid.ca/support/software/amber. Accessed March 2011.

[29] http://www.who.int/csr/disease/meningococcal/en/index.html. Accessed July 2010.

[30] http://www.who.int/mediacentre/factsheets/fs141/en/. Accessed May 2011.

[31] A. K. Abbas, A. H. Lichtman, D. L. Baker, and A. Baker. *Basic Immunology: Functions And Disorders Of The Immune System*. WB Saunders Philadelphia, 2009.

[32] R. Apweiler, A. Bairoch, and C. H. Wu. Protein Sequence Databases. *Current Opinion in Chemical Biology*, 8(1):76 – 80, 2004.

[33] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede. The SWISS-MODEL Workspace: A Web-Based Environment For Protein Structure Homology Modelling. *Bioinformatics*, 22(2):195–201, 2006.

[34] P. T. Beernink and D. M. Granoff. The Modular Architecture Of Meningococcal Factor H-Binding Protein. *Microbiology*, 155(9):2873, 2009.

[35] J. D. Bendtsen, H. Nielsen, G. Heijne, and S. Brunak. Improved Prediction Of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology*, 340(4):783 – 795, 2004.

[36] M. J. Betts and R. B. Russell. *Amino Acid Properties And Consequences Of Substitutions*, pages 289–316. John Wiley & Sons, Ltd, 2003.

[37] C. Brehony, D. J. Wilson, and M. C. J. Maiden. Variation Of The Factor H-Binding Protein Of Neisseria Meningitidis. *Microbiology*, 155(12):4155–4169, 2009.

[38] M. Comanducci, S. Bambini, B. Brunelli, J. Adu-Bobie, B. Arico, B. Capecchi, M. M. Giuliani, V. Masignani, L. Santini, S. Savino, et al. NadA, A Novel Vaccine Candidate Of Neisseria Meningitidis. *Journal of Experimental Medicine*, 195(11):1445, 2002.

[39] S. R. de Còrdoba, J. Esparza-Gordillo, E. G. de Jorge, M. Lopez-Trascasa, and P. Snchez-Corral. The Human Complement Factor H: Functional Roles, Genetic Variations And Disease Associations. *Molecular Immunology*, 41(4):355 – 367, 2004.

[40] J. R. DiPersio. Meningococcal Disease: Prevention And Control Revisited. *Clinical Microbiology Newsletter*, 23(20):155–159, 2001.

[41] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen. Locating Proteins In The Cell Using TargetP, SignalP and Related Tools. *Nat. Protocols*, 2(4):953–971, 04 2007.

[42] L. D. Fletcher, L. Bernfield, V. Barniak, J. E. Farley, A. Howell, M. Knauf, P. Ooi, R. P. Smith, P. Weise, M. Wetherell, et al. Vaccine Potential Of The Neisseria Meningitidis 2086 Lipoprotein. *Infection and immunity*, 72(4):2088, 2004.

[43] M. M. Giuliani, L. Santini, B. Brunelli, A. Biolchi, B. Arico, F. Di Marcello, E. Cartocci, M. Comanducci, V. Masignani, L. Lozzi, et al. The Region Comprising Amino Acids 100 To 255 Of Neisseria Meningitidis Lipoprotein GNA 1870 Elicits Bactericidal Antibodies. *Infection and immunity*, 73(2):1151, 2005.

[44] R. Guerois, J. E. Nielsen, and L. Serrano. Predicting Changes In The Stability Of Proteins And Protein Complexes: A Study Of More Than 1000 Mutations. *Journal of Molecular Biology*, 320(2):369 – 387, 2002.

[45] N. Guex and M. C. Peitsch. SWISS-MODEL And The Swiss-Pdb Viewer: An Environment For Comparative Protein Modeling. *Electrophoresis*, 18(15):2714–2723, 1997.

[46] E. Keedwell and A. Narayanan. *Intelligent Bioinformatics: The Application Of Artificial Intelligence Techniques To Bioinformatics Problems*. Wiley, 2005.

[47] F. Kiefer, K. Arnold, M. Kunzli, L. Bordoli, and T. Schwede. The SWISS-MODEL Repository And Associated Resources. *Nucl. Acids Res.*, 37(suppl_1):D387–392, 2009.

[48] T. J. Kindt, R. A. Goldsby, B. A. Osborne, and J. Kuby. *Kuby Immunology*. WH Freeman, 2007.

[49] P. Kollman. AMBER (Assisted Model Building With Energy Refinement). *University of California, San Francisco, USA*, 2001.

[50] C. Lambert, N. Lèonard, X. De Bolle, and E. Depiereux. ESyPred3D: Prediction Of Proteins 3D Structures. *Bioinformatics*, 18(9):1250–1256, 2002.

[51] Y. Lee, T. Kim, T. Kang, W. Chung, and G. Shin. WSPMaker: A Web Tool For Calculating Selection Pressure In Proteins And Domains Using Window-Sliding. *BMC Bioinformatics*, 9(Suppl 12):S13, 2008.

[52] W. Li and A. Godzik. Cd-Hit: A Fast Program For Clustering And Comparing Large Sets Of Protein Or Nucleotide Sequences. *Bioinformatics*, 22(13):1658, 2006.

[53] W. Li, L. Jaroszewski, and A. Godzik. Clustering Of Highly Homologous Sequences To Reduce The Size Of Large Protein Databases. *Bioinformatics*, 17(3):282, 2001.

[54] W. Li, L. Jaroszewski, and A. Godzik. Tolerating Some Redundancy Significantly Speeds Up Clustering Of Large Protein Databases. *Bioinformatics*, 18(1):77, 2002.

[55] V. Masignani, M. Comanducci, M. M. Giuliani, S. Bambini, J. Adu-Bobie, B. Arico, B. Brunelli, A. Pieri, L. Santini, S. Savino, et al. Vaccination Against Neisseria Meningitidis Using Three Variants Of The Lipoprotein GNA1870. *Journal of Experimental Medicine*, 197(6):789, 2003.

[56] T. Massingham and N. Goldman. Detecting Amino Acid Sites Under Positive Selection And Purifying Selection. *Genetics*, 169(3):1753, 2005.

[57] D. W. Mount. *Bioinformatics: Sequence And Genome Analysis*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2001.

[58] D. Mrozek, B. Maysiak-Mrozek, S. Kozielski, and S. Gòrczyǹska-Kosiorz. Energy Properties Of Protein Structures In The Analysis Of The Human RAB5A Cellular Activity. In K. Cyran, S. Kozielski, J. Peters, U. Stanczyk, and A. Wakulicz-Deja, editors, *Man-Machine Interactions*, volume 59 of *Advances in Soft Computing*, pages 121–131. Springer Berlin / Heidelberg.

[59] E. Murphy, L. Andrew, K. Lee, D. A. Dilts, L. Nunez, P. S. Fink, K. Ambrose, R. Borrow, J. Findlow, M. Taha, et al. Sequence Diversity Of The Factor H Binding Protein Vaccine Candidate In Epidemiologically Relevant Strains Of Serogroup B Neisseria Meningitidis. *Journal of Infectious Diseases*, 200(3):379–389, 2009.

[60] M. Nei. *Molecular Evolutionary Genetics*. Columbia Univ Pr, 1987.

[61] M. Nei and T. Gojobori. Simple Methods For Estimating The Numbers Of Synonymous And Nonsynonymous Nucleotide Substitutions. *Molecular Biology and Evolution*, 3(5):418–426, 1986.

[62] M. Nei and S. Kumar. *Molecular Evolution And Phylogenetics*. Oxford University Press, USA, 2000.

[63] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne. Identification Of Prokaryotic And Eukaryotic Signal Peptides And Prediction Of Their Cleavage Sites. *Protein Engineering*, 10(1):1–6, 1997.

[64] M. Nielsen, C. Lundegaard, O. Lund, and T. N. Petersen. CPHmodels-3.0: A Remote Homology Modeling Using Structure-Guided Sequence Profiles. *Nucleic Acids Research*, 38(suppl 2):W576–W581, 2010.

[65] C. Notredame. Recent Progress In Multiple Sequence Alignment: A Survey. *Pharmacogenomics*, 3(1):131–144, 2002.

[66] World Health Organization. *World Health Statistics 2008. Nonserial Publication.* World Heath Organization, 2008.

[67] M. C. Peitsch. Protein Modeling By E-Mail. *Nat Biotech*, 13(7):658–660, 1995.

[68] M. Pizza, J. Donnelly, and R. Rappuoli. Factor H-Binding Protein, A Unique Meningococcal Vaccine Antigen. *Vaccine*, 26:I46–I48, 2008.

[69] R. A. Sayle and E. J. Milner-White. RASMOL: Biomolecular Graphics For All. *Trends in Biochemical Sciences*, 20(9):374 – 376, 1995.

[70] M. C. Schneider, B. E. Prosser, J. J. E. Caesar, E. Kugelberg, S. Li, Q. Zhang, S. Quoraishi, J. E. Lovett, J. E. Deane, R. B. Sim, et al. Neisseria Meningitidis Recruits Factor H Using Protein Mimicry Of Host Carbohydrates. *Nature*, 458(7240):890–893, 2009.

[71] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.

[72] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The FoldX Web Server: An Online Force Field. *Nucl. Acids Res.*, 33(suppl_2):W382–388, 2005.

[73] J. W. H. Schymkowitz, F. Rousseau, I. C. Martins, J. Ferkinghoff-Borg, F. Stricher, and L. Serrano. Prediction Of Water And Metal Binding Sites And Their Affinities By Using The Fold-X Force Field. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10147–10152, 2005.

[74] F. Stricher, T. Lenaerts, J. Schymkowitz, F. Rousseau, and L. Serrano. FoldX 3.0. In Preparation. 2008.

[75] Y. Suzuki and T. Gojobori. A Method For Detecting Positive Selection At Single Amino Acid Sites. *Molecular Biology and Evolution*, 16(10):1315–1328, 1999.

[76] K. Tamura and J. Dudley. Nei. M., And Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution*, 24:1596–1599.

[77] G. Von Heijne. The Structure Of Signal Peptides From Bacterial Lipoproteins. *Protein engineering*, 2(7):531, 1989.

[78] L. Wall and M. Loukides. *Programming Perl.* O'Reilly & Associates, Inc. Sebastopol, CA, USA, 2000.

[79] I. M. Wallace, G. Blackshields, and D. G. Higgins. Multiple Sequence Alignments. *Current Opinion in Structural Biology*, 15(3):261 – 266, 2005.

[80] J. A. Welsch, S. Ram, O. Koeberling, and D. M. Granoff. Complement-Dependent Synergistic Bactericidal Activity Of Antibodies Against Factor H–Binding Protein, A Sparsely Distributed Meningococcal Vaccine Antigen. *The Journal of infectious diseases*, 197:1053–1061, 2008.

[81] X. Xia and S. Kumar. Codon-Based Detection Of Positive Selection Can Be Biased By Heterogeneous Distribution Of Polar Amino Acids Along Protein Sequences. In *Computational systems bioinformatics: CSB2006 conference proceedings, Stanford CA, 14-18 August 2006*, page 335. Imperial College Pr, 2006.

[82] Z. Yang and J. P. Bielawski. Statistical Methods For Detecting Molecular Adaptation. *Trends in ecology and evolution*, 15(12):496–502, 2000.

[83] Z. Zhang, J. Li, X. Zhao, J. Wang, G. K. Wong, and J. Yu. KaKs_Calculator: Calculating Ka And Ks Through Model Selection And Model Averaging. *Genomics, Proteomics & Bioinformatics*, 4(4):259 – 263, 2006.

[84] P. F. Zipfel, T. Hallström, S. Hammerschmidt, and C. Skerka. The Complement Fitness Factor H: Role In Human Diseases And For Immune Escape Of Pathogens, Like Pneumococci. *Vaccine*, 26(Supplement 8):I67 – I74, 2008.

Appendix A

Scripts

The following scripts were used to identify the conserved, restricted variable, and unrestricted variable regions of the multiple sequence alignment and to determine the amino acid substitution domains for the variable regions (Steps 6 and 11 of the Methodology, Section 4.2).

A.1 Script 01

This script generates a table which depicts how often each symbol appears in each column of the alignment excluding the 'gap' symbols. Columns of the table are delimited by tab characters. Each row of the table begins with a position of the multiple sequence alignment. There follow some number of columns, possibly only 1, where each column consists of an occurrence count and the amino acid code with that occurrence count separated by a space.

```
1.  #!/bin/csh
2.  set DATASET=MSA_seq
3.  set N=263
4.  set INFILE=${DATASET}.msf
5.  set OTFILE=${DATASET}.cnts.nogaps.script.txt
6.  echo -n > !  $OTFILE
7.  foreach I (`jot $N 1 $N 1`)
8.       printf "%d\t" $I >> $OTFILE
9.       seqret -sequence "${INFILE}:*[${I}:${I}]" -outseq stdout -auto | \
10.         grep -v '^>' | grep -v '-' | \
11.         sort | uniq -c | \
12.         tr -s '  ' | tr '\n' '\t' >> $OTFILE
13.      echo >> $OTFILE
14.  end
```

Explanation for Script 01.

- Line 2: set the name of the dataset.
- Line 3: set the variable `'N'` to the value 263, the maximum length of the fHbp sequences after removal of the signal protein part.
- Lines 4 and 5: set the names of the input and output files based on the dataset name.
- Line 6: create the output file, or if it already exists, empty it of previous content.
- Lines 7 through 14: iteratively examine each position of the alignment, with positions ranging from 1 to 263. The `jot` command produces the sequence of numbers from 1 to 263 in increments of 1. The loop index variable is `I`.
- Line 8: output the position number in the sequence.
- Lines 9 through 12: this set of lines constitutes a single UNIX command pipeline. The first step is to obtain the amino acid in position `I` for all the sequences where a multi-FASTA file with 190 single-character sequences is generated. The next step is to delete all the header records (lines starting with '>') from the multi-FASTA file, generating 190 one-character lines of raw sequence information and then remove all occurrences of the 'gap' character. Count the number of each character. Finally produce an output line that contains tab-delimited

columns. Each column consists of an occurrence count and the character with that count separated by a space.

- Line 9: have the EMBOSS `seqret` command produce on the standard output, a multi-FASTA file with 190 single-character sequences in it.
- Line 10: the two `grep` commands here act as filters. UNIX command `grep -v` inverts the sense of matching. Command `grep -v '^>'` removes any line which starts with the '>' symbol and `grep -v '-'` excludes the 'gap' symbol.
- Line 11: sort the set of amino acids using the UNIX `sort` command, and then count the number of occurrences of each amino acid using the command `uniq -c`. The latter command will produce a table with one line for each amino acid. On the line will be a count and an amino acid separated by space characters.
- Line 12: UNIX command `tr -s ' '` replaces the repeated occurrences of the character ' ' with its single occurrence. The second command on line 12 replaces each newline character by a tab character. The purpose is so that all the output related to a particular multiple sequence alignment position is contained in a single line to be output.
- Line 13: terminate the line of output from line 12 with a newline character.
- Line 14: terminate the *for loop*.

A.2 Script 02

This script generates a table which depicts how often each symbol appears in each column of the alignment including the 'gap' symbols. Columns of the table are delimited by tab characters. Each row of the table begins with a position of the multiple sequence alignment. There follow some number of columns, possibly only 1, where each column consists of an occurrence count and the amino acid code or 'gap' symbol, with that occurrence count. The latter two fields in each column are separated by a space.

```
1.  #!/bin/csh
2.  set DATASET=MSA_seq
3.  set N=263
4.  set INFILE=${DATASET}.msf
5.  set OTFILE=${DATASET}.cnts.gaps.script.txt
6.  echo -n > !  $OTFILE
7.  foreach I (`jot $N 1 $N 1`)
8.       printf "%d\t" $I >> $OTFILE
9.       seqret -sequence "${INFILE}:*[${I}:${I}]" -outseq stdout -auto | \
10.         grep -v '^>' | \
11.         sort | uniq -c | \
12.         tr -s '  ' | tr '\n' '\t' >> $OTFILE
13.      echo >> $OTFILE
14.  end
```

Explanation for Script 02.

- Line 2: set the name of the dataset.
- Line 3: set the variable `'N'` to the value 263, the maximum length of the fHbp sequences after removal of the signal protein part.
- Lines 4 and 5: set the names of the input and output files based on the dataset name.

- Line 6: create the output file, or if it already exists, empty it of previous content.
- Lines 7 through 14: iteratively examine each position of the alignment, with positions ranging from 1 to 263. The `jot` command produces the sequence of numbers from 1 to 263 in increments of 1. The loop index variable is `I`.
- Line 8: output the position number in the sequence.
- Lines 9 through 12: this set of lines constitutes a single UNIX command pipeline. The first step is to obtain the amino acid or the 'gap' symbol in position `I` for all the sequences where a multi-FASTA file with 190 single-character sequences is generated. The next step is to delete all the header records (lines starting with '>') from the multi-FASTA file, generating 190 one-character lines of raw sequence information. Count the number of each character. Finally produce an output line that contains tab-delimited columns. Each column consists of an occurrence count and the character with that count separated by a space.
- Line 9: have the EMBOSS `seqret` command produce on the standard output, a multi-FASTA file with 190 single-character sequences in it.
- Line 10: the `grep` command here act as a filter. UNIX command `grep -v` inverts the sense of matching. Command `grep -v` '∧>' removes any line which starts with the '>' symbol.
- Line 11: sort the set of amino acids and 'gap' symbols using the UNIX `sort` command, and then count the number of occurrences of each amino acid or 'gap' using the command `uniq -c`. The latter command will produce a table with one line for each amino acid or 'gap' symbol. On the line will be a count and an amino acid or 'gap' symbol separated by space characters.
- Line 12: UNIX command `tr -s ' '` replaces the repeated occurrences of the character ' ' with its single occurrence. The second command in line 12 replaces each newline character by a tab character. The purpose is so that all the output related to a particular multiple sequence alignment position is contained in a single line to be output.
- Line 13: terminate the line of output from line 12 with a newline character.
- Line 14: terminate the *for loop*.

A.3 Script 03

This script determines which columns of an alignment have at least one 'gap' symbol and produces a list of those columns, one column number per line.

```
1.  #!/bin/csh
2.  set DATASET=MSA_seq
3.  set N=263
4.  set INFILE=${DATASET}.msf
5.  set OTFILE=${DATASET}.cols_with_gaps.script.txt
6.  echo -n > !  $OTFILE
7.  foreach I (`jot $N 1 $N 1`)
8.       seqret -sequence "${INFILE}:*[${I}:${I}]" -outseq stdout -auto | \
9.         grep -v '∧>' | grep -e '-' > /dev/null
10.      if ( $status == 0 ) then
11.          echo $I >> $OTFILE
12.      endif
13.  end
```

Explanation for Script 03.

- Line 2: set the name of the dataset.
- Line 3: set the variable `'N'` to the value 263, the maximum length of the fHbp sequences after removal of the signal protein part.
- Lines 4 and 5: set the names of the input and output files based on the dataset name.
- Line 6: create the output file, or if it already exists, empty it of previous content.
- Lines 7 through 13: iteratively examine each position of the alignment, with positions ranging from 1 to 263. The `jot` command produces the sequence of numbers from 1 to 263 in increments of 1. The loop index variable is `I`.
- Lines 8 and 9: this set of lines constitutes a single UNIX command pipeline. The exit status of the pipeline will be "success" if the current column position of the alignment (stored in variable `I`) contains any 'gap' symbols.
- Line 8: have the EMBOSS `seqret` command produce on the standard output, a multi-FASTA file with 190 single-character sequences in it.
- Line 9: the first `grep` command here act as a filter. UNIX command `grep -v` inverts the sense of matching. Command `grep -v` '∧>' removes any line which starts with the '>' symbol and `grep -e` '`-`' specifies the matching of the pattern beginning with '`-`' symbol. The `/dev/null` used here is a special file that discards all data written to it and in this case '`-`' is written to it, whenever it is found.
- Lines 10 through 12: constitutes the *if* condition. The *if* condition checks the exit status returned by the last command, which in this case is the pipeline in lines 8 and 9. The csh variable `status` returns a status of 1 for an unsuccessful exit of the pipeline and returns a 0 for a successful exit of the pipeline.
- Line 11: command is executed only when the *if* condition of Line 10 is true. Output consists of the position number in the sequence alignment with at least one 'gap' symbol.
- Line 13: terminate the *for loop*.

A.4 Script 04

This script determines the frequency distribution for the counts of the different amino acids across the columns of the alignment, with 'gap' symbols excluded and involves the use of an intermediate file. The intermediate result generates a file where each row represents each column position in the multiple sequence alignment followed by the count of the number of unique amino acids occurring in that position delimited by a tab character. For instance, suppose the first row has the output `1` followed by `1` delimited by a tab character, the second row has `2` followed by `1`, the third row has `3` followed by `2` and so on. This means that the first column position in the sequence alignment has 1 type of amino acid, the second position has 1 type too, the third position have 2 unique amino acids, etc. The final phase of this script reads the intermediate result as input and outputs a file which consists of two columns separated by a space character. The first column is the frequency of the counts; that is, how often each count appears and the second column is the count itself. For instance, suppose the first row in the final output has `122` followed by `1` separated by a space character. This means that in our alignment ignoring the occurrences of the 'gap' symbols, there were 122 positions with only one type of amino acid.

```
1.  #!/bin/csh
2.  set DATASET=MSA_seq
3.  set N=263
4.  set INFILE=${DATASET}.msf
5.  set OTFILE=${DATASET}.cnts_residues_each_column.nogaps.script.txt
6.  echo -n > !  $OTFILE
7.  foreach I (`jot $N 1 $N 1`)
8.       printf "%d\t" $I >> $OTFILE
9.       seqret -sequence "${INFILE}:*[${I}:${I}]" -outseq stdout -auto | \
10.         grep -v `^>' | grep -v `-' | \
11.         sort -u | wc -l | \
12.         tr -d "  " >> $OTFILE
13.  end
14.  set INFILE=${DATASET}.cnts_residues_each_column.nogaps.script.txt
15.  set OTFILE=${DATASET}.freq.nogaps.script.txt
16.  cut -f 2 $INFILE | sort | uniq -c > !  $OTFILE
```

Explanation for Script 04.

This script has two parts. The first part outputs the count of the number of unique amino acids for each column position in the alignment excluding the 'gap' symbols. This output becomes the input for the next part of this script which determines a frequency distribution for the intermediate result generated in the first part. Note that the 'gap' symbols are excluded in this script. Lines 1 to 13 form the first phase of this script and lines 14 to 16 the second phase.

- Line 2: set the name of the dataset.
- Line 3: set the variable `'N'` to the value 263, the maximum length of the fHbp sequences after removal of the signal protein part.
- Lines 4 and 5: set the names of the input and output files based on the dataset name.
- Line 6: create the output file, or if it already exists, empty it of previous content.
- Lines 7 through 13: iteratively examine each position of the alignment, with positions ranging from 1 to 263. For each position, output to the intermediate output file a line which contains two tab-separated values: the position number and a count of the number of unique amino acids occurring at that position. The `jot` command produces the sequence of numbers from 1 to 263 in increments of 1. The loop index variable is `I`.
- Line 8: output the position number in the sequence.
- Lines 9 through 12: this set of lines constitutes a single UNIX command pipeline. The pipeline produces a count of how many different amino acids occur at this position (in the alignment).
- Line 9: have the EMBOSS `seqret` command produce on the standard output a multi-FASTA file with 190 single-character sequences in it.
- Line 10: the two `grep` commands here act as filters. UNIX command `grep -v` inverts the sense of matching. Command `grep -v` '^>' removes any line which starts with the '>' symbol and `grep -v` '-' excludes the 'gap' symbol.
- Line 11: UNIX command `sort -u` sorts the list of amino acids and suppresses all but one occurrence of matching keys. The command `wc -l` counts the lines generated; i.e. it counts the number of unique amino acids. Note that this output from `wc` will be terminated by a newline character.
- Line 12: UNIX command `tr -d ' '` removes any space characters surrounding the count, and outputs the bare count to the intermediate output file.

- Line 13: terminate the *for loop*.
- Line 14: set the output file generated in the first phase as the input file for the second phase.
- Line 15: set the name of the output file based on the dataset name.
- Line 16: UNIX command `cut -f 2 $INFILE` extracts just the second column from the input file, which contains the count of the different amino acids occurring for each particular position. The UNIX command `sort` sorts these counts and `uniq -c` determines the frequency of the counts; that is, the `uniq` command will determine how often each count occurs. It will then produce a two column table, with the columns separated by a space, with a frequency in the first column and a count in the second column and forwards it to the output file of the second phase.

A.5 Script 05

This script determines the frequency distribution of the counts of the different amino acids, including the occurrences of the 'gap' symbols across the columns of the alignment, and involves the use of an intermediate file. The intermediate result generates a file where each row represents each column position in the multiple sequence alignment followed by the count of the number of unique amino acids or 'gaps' occurring in that position delimited by a tab character. For instance, suppose the first row has the output `1` followed by `1` delimited by a tab character, the second row has `2` followed by `1`, the third row has `3` followed by `2` and so on. This means that the first column position in the sequence alignment has 1 type of amino acid or just 'gaps', the second position has 1 type too, the third position have 2 unique amino acids or one type of amino acid and 'gaps', etc. The final phase of this script reads the intermediate result as input and outputs a file which consists of two columns separated by a space character. The first column is the frequency of the counts; that is, how often each count appears and the second column is the count itself. For instance, suppose the first row in the final output has `115` followed by `1` separated by a space character. This means that in our alignment including the occurrences of the 'gap' symbols, there were 115 positions with only one type of amino acid or 'gaps'.

```
1.  #!/bin/csh
2.  set DATASET=MSA_seq
3.  set N=263
4.  set INFILE=${DATASET}.msf
5.  set OTFILE=${DATASET}.cnts_residues_each_column.gaps.script.txt
6.  echo -n > !  $OTFILE
7.  foreach I (`jot $N 1 $N 1`)
8.       printf "%d\t" $I >> $OTFILE
9.       seqret -sequence "${INFILE}:*[${I}:${I}]" -outseq stdout -auto | \
10.         grep -v '^>' | \
11.         sort -u | wc -l | \
12.         tr -d "  " >> $OTFILE
13.  end
14.  set INFILE=${DATASET}.cnts_residues_each_column.gaps.script.txt
15.  set OTFILE=${DATASET}.freq.gaps.script.txt
16.  cut -f 2 $INFILE | sort | uniq -c > !  $OTFILE
```

Explanation for Script 05.

This script has two parts. The first part outputs the count of the number of unique amino acids including the 'gap' symbols for each column position in the alignment. This output becomes the input for the next part of this script which determines a frequency distribution for the intermediate result generated in the first part. Lines 1 to 13 form the first phase of this script and lines 14 to 16 the second phase.

- Line 2: set the name of the dataset.
- Line 3: set the variable `'N'` to the value 263, the maximum length of the fHbp sequences after removal of the signal protein part.
- Lines 4 and 5: set the names of the input and output files based on the dataset name.
- Line 6: create the output file, or if it already exists, empty it of previous content.
- Lines 7 through 13: iteratively examine each position of the alignment, with positions ranging from 1 to 263. For each position, output to the intermediate output file a line which contains two tab-separated values: the position number and a count of the number of unique amino acids and 'gaps' occurring at that position. The `jot` command produces the sequence of numbers from 1 to 263 in increments of 1. The loop index variable is `I`.
- Lines 9 through 12: this set of lines constitutes a single UNIX command pipeline. The pipeline produces a count of how many different amino acids and 'gaps' occur at this position (in the alignment).
- Line 9: have the EMBOSS `seqret` command produce on the standard output a multi-FASTA file with 190 single-character sequences in it.
- Line 10: the `grep` command here act as a filter. UNIX command `grep -v` inverts the sense of matching. Command `grep -v` '$\wedge>$' removes any line which starts with the '$>$' symbol.
- Line 11: UNIX command `sort -u` sorts the list of amino acids and 'gaps' and suppresses all but one occurrence of matching keys. The command `wc -l` counts the lines generated; i.e. it counts the number of unique amino acids or the occurrences of 'gap' symbols. Note that this output from `wc` will be terminated by a newline character.
- Line 12: UNIX command `tr -d ' '` removes any space characters surrounding the count, and outputs the bare count to the intermediate output file.
- Line 13: terminate the *for loop*.
- Line 14: set the output file generated in the first phase as the input file for the second phase.
- Line 15: set the name of the output file based on the dataset name.
- Line 16: UNIX command `cut -f 2 $INFILE` extracts just the second column from the input file, which contains the count of the different amino acids including the 'gaps' occurring for each particular position. The UNIX command `sort` sorts these counts and `uniq -c` determines the frequency of the counts; that is, the `uniq` command will determine how often each count occurs. It will then produce a two column table, with the columns separated by a space, with a frequency in the first column and a count in the second column and forwards it to the output file of the second phase.

Appendix B

Tables of Results

Table B.1: Frequency of different amino acids in each position of the 190 fHbp sequences with 'gap' symbol excluded. The results of this table were generated using Script A.1 of Appendix A.

Position	190 fHbp sequences
1	190C
2	190S
3	190S
4	190G
5	134G 2S
6	58G
7	58G
8	57S
9	56G
10	100G 11S
11	190G
12	190G
13	3I 187V
14	183A 7T
15	189A 1V
16	190D
17	190I
18	190G
19	138A 48T 4V
20	187G 1R 2V
21	190L
22	190A
23	189D 1Y
24	190A
25	190L
26	190T
27	173A 16T 1V
28	2L 188P
29	2F 186L 2P
30	190D
31	190H
32	190K
33	189D 1N
34	190K
35	144G 46S
36	190L
37	53K 133Q 4R
38	190S
39	190L
40	1A 1I 18M 170T
41	190L
42	136D 53E 1N
43	53D 137Q
44	190S
45	53I 137V
46	35P 137R 18S
47	136K 53Q 1R
48	3K 187N
49	137E 53G
50	137K 53T

Continued on next page

Table B.1 – continued from previous page

Position	190 fHbp sequences
51	190L
52	137K 53T
53	190L
54	133A 57S
55	190A
56	190Q
57	190G
58	189A 1V
59	190E
60	4E 177K 9R
61	2I 188T
62	50F 140Y
63	140G 50K
64	42A 140N 8V
65	190G
66	187D 2G 1N
67	50K
68	50D
69	50N
70	1N 2R 187S
71	190L
72	3D 186N 1S
73	190T
74	189G 1S
75	190K
76	190L
77	190K
78	190N
79	190D
80	190K
81	64I 126V
82	190S
83	190R
84	190F
85	190D
86	190F
87	148I 42V
88	4H 42Q 144R
89	42K 148Q
90	190I
91	189E 1R
92	1S 189V
93	188D 2N
94	186G 4R
95	14K 176Q
96	142L 48T
97	190I
98	190T
99	190L
100	48A 142E
101	2N 1R 187S
102	190G
103	190E
104	190F
105	190Q
106	83I 107V
107	190Y
108	190K
109	190Q
110	68D 1G 15N 106S
111	186H 4Y
112	190S

Continued on next page

Table B.1 – continued from previous page

Position	190 fHbp sequences
113	190A
114	106L 84V
115	106T 84V
116	190A
117	18F 172L
118	190Q
119	82I 108T
120	190E
121	84K 106Q
122	20E 100I 70V
123	84N 106Q
124	106D 84N
125	2L 99P 89S
126	84D 106E
127	33D 73H 84K
128	77I 106S 7T
129	1A 84D 6E 94G 5R
130	104K 86S
131	84L 106M
132	84I 106V
133	105A 84N 1V
134	106K 83Q 1R
135	190R
136	78Q 28R 84S
137	190F
138	16K 84L 90R
139	106I 84V
140	106G 84S
141	106D 83G 1S
142	106I 84L
143	100A 84G 6V
144	190G
145	190E
146	190H
147	1I 189T
148	84A 106S
149	190F
150	102D 4G 84N
151	106K 84Q
152	190L
153	1H 188P 1R
154	1D 49E 1G 55K
155	45D 112G 30S 3V
156	13D 143G 3S 31V
157	84K 28M 51R 27S
158	188A 1S 1V
159	84E 106T
160	190Y
161	84H 106R
162	190G
163	84K 106T
164	190A
165	187F 3L
166	89G 101S
167	190S
168	190D
169	190D
170	151A 38P 1T
171	144G 38N 1R 7S
172	190G
173	1E 150K 39R
174	190L

Continued on next page

Table B.1 – continued from previous page

Position	190 fHbp sequences
175	38H 15I 137T
176	190Y
177	33S 157T
178	190I
179	190D
180	190F
181	151A 38T 1V
182	146A 33K 5N 1S 2T 3V
183	190K
184	190Q
185	190G
186	1C 131H 12N 46Y
187	190G
188	1G 151K 38R
189	190I
190	190E
191	190H
192	190L
193	190K
194	105S 85T
195	1L 189P
196	190E
197	107L 83Q
198	190N
199	190V
200	74D 115E 1N
201	190L
202	188A 2V
203	99A 53S 29T 9V
204	182A 8S
205	48D 85E 1N 56Y
206	105I 85L
207	1E 189K
208	85A 104P 1Q
209	190D
210	169E 9G 12K
211	190K
212	53H 52R 85S
213	188H 2Y
214	190A
215	190V
216	190I
217	85L 105S
218	190G
219	85D 1F 104S
220	85T 105V
221	1H 105L 84R
222	190Y
223	1D 84G 105N
224	28G 1H 104Q 57S
225	40A 64D 85E 1N
226	190E
227	190K
228	189G 1S
229	105S 85T
230	190Y
231	85H 105S
232	190L
233	85A 105G
234	105I 85L
235	190F
236	189G 1S

Continued on next page

Table B.1 – continued from previous page

Position	190 fHbp sequences
237	85D 10E 94G 1R
238	1E 50K 54Q 85R
239	190A
240	190Q
241	190E
242	86I 104V
243	190A
244	190G
245	190S
246	190A
247	105E 85T
248	190V
249	57E 133K
250	85I 105T
251	64A 41G 44R 41V
252	85E 105N
253	105G 85K
254	105I 85V
255	127H 7Q 56R
256	85E 105H
257	190I
258	1D 183G 6S
259	85I 105L
260	190A
261	105A 85G
262	190K
263	190Q

Table B.2: Frequency of different amino acids and the 'gap' symbol in each position of the 190 fHbp sequences and the two sets of new variants, Set A and Set B. The results of this table were generated using Script A.2 of Appendix A.

Position	190 fHbp sequences	Set A (200 new variants)	Set B (100 new variants)
1	190C	200C	100C
2	190S	200S	100S
3	190S	200S	100S
4	190G	200G	100G
5	54 – 134G 2S	50 – 148G 2S	30 – 70G
6	132 – 58G	135 – 65G	64 – 36G
7	132 – 58G	144 – 56G	73 – 27G
8	133 – 57S	141 – 59S	72 – 28S
9	134 – 56G	152 – 48G	69 – 31G
10	79 – 100G 11S	72 – 109G 19S	32 – 62G 6S
11	190G	200G	100G
12	190G	200G	100G
13	3I 187V	6I 194V	4I 96V
14	183A 7T	194A 6T	97A 3T
15	189A 1V	197A 3V	98A 2V
16	190D	200D	100D
17	190I	200I	100I
18	190G	200G	100G
19	138A 48T 4V	152A 42T 6V	78A 21T 1V
20	187G 1R 2V	194G 1R 5V	99G 1R
21	190L	200L	100L
22	190A	200A	100A
23	189D 1Y	198D 2Y	100D
24	190A	200A	100A
25	190L	200L	100L
26	190T	200T	100T

Continued on next page

Table B.2 – continued from previous page

Position	190 fHbp sequences	Set A (200 new variants)	Set B (100 new variants)
27	173A 16T 1V	179A 16T 5V	92A 8T
28	2L 188P	1L 199P	3L 97P
29	2F 186L 2P	4F 193L 3P	97L 3P
30	190D	200D	100D
31	190H	200H	100H
32	190K	200K	100K
33	189D 1N	199D 1N	100D
34	190K	200K	100K
35	144G 46S	144G 56S	78G 22S
36	190L	200L	100L
37	53K 133Q 4R	57K 142Q 1R	29K 70Q 1R
38	190S	200S	100S
39	190L	200L	100L
40	1A 1I 18M 170T	1A 3I 19M 177T	1I 8M 91T
41	190L	200L	100L
42	136D 53E 1N	142D 57E 1N	70D 29E 1N
43	53D 137Q	57D 143Q	29D 71Q
44	190S	200S	100S
45	53I 137V	57I 143V	29I 71V
46	35P 137R 18S	36P 143R 21S	16P 71R 13S
47	136K 53Q 1R	142K 57Q 1R	70K 29Q 1R
48	3K 187N	10K 190N	3K 97N
49	137E 53G	143E 57G	71E 29G
50	137K 53T	143K 57T	71K 29T
51	190L	200L	100L
52	137K 53T	143K 57T	71K 29T
53	190L	200L	100L
54	133A 57S	142A 58S	70A 30S
55	190A	200A	100A
56	190Q	200Q	100Q
57	190G	200G	100G
58	189A 1V	199A 1V	99A 1V
59	190E	200E	100E
60	4E 177K 9R	2E 191K 7R	1E 96K 3R
61	2I 188T	1I 199T	3I 97T
62	50F 140Y	52F 148Y	33F 67Y
63	140G 50K	148G 52K	67G 33K
64	42A 140N 8V	43A 148N 9V	32A 67N 1V
65	190G	200G	100G
66	187D 2G 1N	194D 2G 4N	96D 1G 3N
67	140 – 50K	148 – 52K	67 – 33K
68	140 – 50D	148 – 52D	67 – 33D
69	140 – 50N	148 – 52N	67 – 33N
70	1N 2R 187S	2N 3R 195S	1N 1R 98S
71	190L	200L	100L
72	3D 186N 1S	2D 193N 5S	1D 97N 2S
73	190T	200T	100T
74	189G 1S	197G 3S	99G 1S
75	190K	200K	100K
76	190L	200L	100L
77	190K	200K	100K
78	190N	200N	100N
79	190D	200D	100D
80	190K	200K	100K
81	64I 126V	78I 122V	30I 70V
82	190S	200S	100S
83	190R	200R	100R
84	190F	200F	100F
85	190D	200D	100D
86	190F	200F	100F
87	148I 42V	153I 47V	77I 23V
88	4H 42Q 144R	3H 47Q 150R	1H 23Q 76R

Continued on next page

Table B.2 – continued from previous page

Position	190 fHbp sequences	Set A (200 new variants)	Set B (100 new variants)
89	42K 148Q	47K 153Q	23K 77Q
90	190I	200I	100I
91	189E 1R	199E 1R	99E 1R
92	1S 189V	2S 198V	100V
93	188D 2N	199D 1N	100D
94	186G 4R	193G 7R	96G 4R
95	14K 176Q	15K 185Q	4K 96Q
96	142L 48T	151L 49T	72L 28T
97	190I	200I	100I
98	190T	200T	100T
99	190L	200L	100L
100	48A 142E	49A 151E	28A 72E
101	2N 1R 187S	2N 2R 196S	1R 99S
102	190G	200G	100G
103	190E	200E	100E
104	190F	200F	100F
105	190Q	200Q	100Q
106	83I 107V	83I 117V	36I 64V
107	190Y	200Y	100Y
108	190K	200K	100K
109	190Q	200Q	100Q
110	68D 1G 15N 106S	65D 1G 15N 119S	33D 7N 60S
111	186H 4Y	198H 2Y	99H 1Y
112	190S	200S	100S
113	190A	200A	100A
114	106L 84V	119L 81V	60L 40V
115	106T 84V	119T 81V	60T 40V
116	190A	200A	100A
117	18F 172L	16F 184L	9F 91L
118	190Q	200Q	100Q
119	82I 108T	91I 109T	48I 52T
120	190E	200E	100E
121	84K 106Q	81K 119Q	40K 60Q
122	20E 100I 70V	17E 105I 78V	7E 59I 34V
123	84N 106Q	81N 119Q	40N 60Q
124	106D 84N	119D 81N	60D 40N
125	2L 99P 89S	2L 98P 100S	1L 45P 54S
126	84D 106E	81D 119E	40D 60E
127	33D 73H 84K	39D 80H 81K	12D 48H 40K
128	77I 106S 7T	74I 119S 7T	37I 60S 3T
129	1A 84D 6E 94G 5R	81D 3E 104G 12R	40D 2E 57G 1R
130	104K 86S	108K 92S	58K 42S
131	84L 106M	81L 119M	40L 60M
132	84I 106V	81I 119V	40I 60V
133	105A 84N 1V	119A 81N	60A 40N
134	106K 83Q 1R	119K 80Q 1R	60K 40Q
135	190R	200R	100R
136	78Q 28R 84S	88Q 31R 81S	49Q 11R 40S
137	190F	200F	100F
138	16K 84L 90R	19K 81L 100R	6K 40L 54R
139	106I 84V	119I 81V	60I 40V
140	106G 84S	119G 81S	60G 40S
141	106D 83G 1S	119D 80G 1S	60D 40G
142	106I 84L	119I 81L	60I 40L
143	100A 84G 6V	112A 81G 7V	58A 40G 2V
144	190G	200G	100G
145	190E	200E	100E
146	190H	200H	100H
147	1I 189T	1I 199T	100T
148	84A 106S	81A 119S	40A 60S
149	190F	200F	100F
150	102D 4G 84N	115D 4G 81N	60D 40N

Continued on next page

Table B.2 – continued from previous page

Position	190 fHbp sequences	Set A (200 new variants)	Set B (100 new variants)
151	106K 84Q	119K 81Q	60K 40Q
152	190L	200L	100L
153	1H 188P 1R	3H 196P 1R	100P
154	84 – 1D 49E 1G 55K	81 – 4D 58E 57K	40 – 23E 37K
155	45D 112G 30S 3V	41D 115G 39S 5V	29D 60G 10S 1V
156	13D 143G 3S 31V	13D 154G 2S 31V	7D 70G 1S 22V
157	84K 28M 51R 27S	81K 36M 53R 30S	43K 15M 31R 11S
158	188A 1S 1V	195A 2S 3V	98A 1S 1V
159	84E 106T	81E 119T	43E 57T
160	190Y	200Y	100Y
161	84H 106R	81H 119R	43H 57R
162	190G	200G	100G
163	84K 106T	81K 119T	40K 60T
164	190A	200A	100A
165	187F 3L	196F 4L	96F 4L
166	89G 101S	100G 100S	43G 57S
167	190S	200S	100S
168	190D	200D	100D
169	190D	200D	100D
170	151A 38P 1T	163A 35P 2T	85A 15P
171	144G 38N 1R 7S	160G 35N 2R 3S	79G 15N 6S
172	190G	200G	100G
173	1E 150K 39R	163K 37R	4E 81K 15R
174	190L	200L	100L
175	38H 15I 137T	35H 12I 153T	15H 12I 73T
176	190Y	200Y	100Y
177	33S 157T	36S 164T	21S 79T
178	190I	200I	100I
179	190D	200D	100D
180	190F	200F	100F
181	151A 38T 1V	163A 35T 2V	79A 21T
182	146A 33K 5N 1S 2T 3V	162A 28K 7N 2S 1V	76A 17K 4N 3T
183	190K	200K	100K
184	190Q	200Q	100Q
185	190G	200G	100G
186	1C 131H 12N 46Y	4C 129H 13N 54Y	1C 65H 9N 25Y
187	190G	200G	100G
188	1G 151K 38R	2G 163K 35R	1G 79K 20R
189	190I	200I	100I
190	190E	200E	100E
191	190H	200H	100H
192	190L	200L	100L
193	190K	200K	100K
194	105S 85T	118S 82T	55S 45T
195	1L 189P	1L 199P	1L 99P
196	190E	200E	100E
197	107L 83Q	116L 84Q	51L 49Q
198	190N	200N	100N
199	190V	200V	100V
200	74D 115E 1N	74D 123E 3N	31D 69E
201	190L	200L	100L
202	188A 2V	198A 2V	98A 2V
203	99A 53S 29T 9V	117AV57SV20T 6V	54A 24S 18T 4V
204	182A 8S	193A 7S	97A 3S
205	48D 85E 1N 56Y	62D 82E 2N 54Y	26D 45E 1N 28Y
206	105I 85L	118I 82L	55I 45L
207	1E 189K	200K	2E 98K
208	85A 104P 1Q	82A 116P 2Q	45A 54P 1Q
209	190D	200D	100D
210	169E 9G 12K	181E 6G 13K	85E 7G 8K
211	190K	200K	100K
212	53H 52R 85S	51H 67R 82S	27H 28R 45S

Continued on next page

Table B.2 – continued from previous page

Position	190 fHbp sequences	Set A (200 new variants)	Set B (100 new variants)
213	188H 2Y	196H 4Y	99H 1Y
214	190A	200A	100A
215	190V	200V	100V
216	190I	200I	100I
217	85L 105S	82L 118S	45L 55S
218	190G	200G	100G
219	85D 1F 104S	82D 2F 116S	45D 1F 54S
220	85T 105V	82T 118V	45T 55V
221	1H 105L 84R	3H 118L 79R	2H 55L 43R
222	190Y	200Y	100Y
223	1D 84G 105N	3D 79G 118N	3D 45G 52N
224	28G 1H 104Q 57S	34G 2H 116Q 48S	18G 1H 51Q 30S
225	40A 64D 85E 1N	52A 64D 82E 2N	23A 28D 48E 1N
226	190E	200E	100E
227	190K	200K	100K
228	189G 1S	198G 2S	98G 2S
229	105S 85T	118S 82T	55S 45T
230	190Y	200Y	100Y
231	85H 105S	82H 118S	45H 55S
232	190L	200L	100L
233	85A 105G	82A 118G	45A 55G
234	105I 85L	118I 82L	55I 45L
235	190F	200F	100F
236	189G 1S	200G	100G
237	85D 10E 94G 1R	82D 13E 103G 2R	45D 5E 49G 1R
238	1E 50K 54Q 85R	2E 65K 51Q 82R	1E 27K 27Q 45R
239	190A	200A	100A
240	190Q	200Q	100Q
241	190E	200E	100E
242	86I 104V	95I 105V	48I 52V
243	190A	200A	100A
244	190G	200G	100G
245	190S	200S	100S
246	190A	200A	100A
247	105E 85T	118E 82T	55E 45T
248	190V	200V	100V
249	57E 133K	60E 140K	35E 65K
250	85I 105T	82I 118T	45I 55T
251	64A 41G 44R 41V	66A 48G 34R 52V	35A 23G 22R 20V
252	85E 105N	82E 118N	45E 55N
253	105G 85K	118G 82K	53G 47K
254	105I 85V	118I 82V	53I 47V
255	127H 7Q 56R	125H 10Q 65R	64H 4Q 32R
256	85E 105H	82E 118H	47E 53H
257	190I	200I	100I
258	1D 183G 6S	1D 196G 3S	96G 4S
259	85I 105L	82I 118L	45I 55L
260	190A	200A	100A
261	105A 85G	118A 82G	55A 45G
262	190K	200K	100K
263	190Q	200Q	100Q

Table B.3: Positions that had (at least one) 'gap' symbol in the 190 fHbp sequences and the two sets of new variants, Set A and Set B. The results in this table were generated using Script A.3 of Appendix A. The table shows that there were 10 positions that had at least one 'gap' symbol.

190 fHbp sequences	Set A (200 new variants)	Set B (100 new variants)
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9
10	10	10
67	67	67
68	68	68
69	69	69
154	154	154

Table B.4: Frequency distribution for the counts of the different amino acids across the positions of the alignment of the 190 fHbp sequences, with 'gap' symbols excluded. For example, only 1 position had 5 different symbols occuring there. The results in this table were generated using Script A.4 in Appendix A.

Frequency	Number of symbols ('gap' excluded)
122	1
82	2
42	3
15	4
1	5
1	6

Table B.5: Frequency distribution for the counts of the different amino acids across the positions of the alignment of the 190 fHbp sequences, with 'gap' symbols included. The results in this table were generated using Script A.5 in Appendix A.

Frequency	Number of symbols ('gap' included)
115	1
87	2
44	3
14	4
2	5
1	6

Table B.6: Codon-based calculation of the K_A/K_S ratio using the sliding window approach. Positions where the K_A/K_S ratios are more than or equal to 0.9 are highlighted with red colour.

Window of the nucleotide positions	**Corresponding amino acid positions**	K_A	K_A/K_S
1-18	1-6	0.002	0.006
4-21	2-7	0.003	0.009
7-24	3-8	0.004	0.012
10-27	4-9	0.002	0.006
13-30	5-10	0.011	0.032
16-33	6-11	0.013	0.038
19-36	7-12	0.009	0.026
22-39	8-13	0.011	0.032
25-42	9-14	0.018	0.052
28-45	10-15	0.016	0.047
31-48	11-16	0.010	0.029
34-51	12-17	0.010	0.029
37-54	13-18	0.010	0.029
40-57	14-19	0.044	0.128
43-60	15-20	0.039	0.114
46-63	16-21	0.038	0.111
49-66	17-22	0.038	0.111
52-69	18-23	0.042	0.122
55-72	19-24	0.042	0.122
58-75	20-25	0.004	0.012
61-78	21-26	0.001	0.003
64-81	22-27	0.017	0.050
67-84	23-28	0.020	0.058
70-87	24-29	0.024	0.070
73-90	25-30	0.024	0.070
76-93	26-31	0.015	0.044
79-96	27-32	0.022	0.064
82-99	28-33	0.007	0.020
85-102	29-34	0.005	0.015
88-105	30-35	0.035	0.102
91-108	31-36	0.038	0.111
94-111	32-37	0.069	0.201
97-114	33-38	0.046	0.134
100-117	34-39	0.055	0.160
103-120	35-40	0.079	0.230
106-123	36-41	0.036	0.105
109-126	37-42	0.032	0.093
112-129	38-43	0.026	0.076
115-132	39-44	0.026	0.076
118-135	40-45	0.025	0.073
121-138	41-46	0.006	0.017
124-141	42-47	0.006	0.017
127-144	43-48	0.258	0.752
130-147	44-49	0.110	0.321
133-150	45-50	0.012	0.035
136-153	46-51	0.160	0.466
139-156	47-52	0.124	0.362
142-159	48-53	0.100	0.292
145-162	49-54	0.202	0.589
148-165	50-55	0.142	0.414
151-168	51-56	0.087	0.254
154-171	52-57	0.089	0.259
157-174	53-58	0.036	0.105
160-177	54-59	0.033	0.096
163-180	55-60	0.013	0.038
166-183	56-61	0.015	0.044
169-186	57-62	0.044	0.128
172-189	58-63	0.138	0.402

Continued on next page

Table B.6 – continued from previous page

Window of the nucleotide positions	Corresponding amino acid positions	K_A	K_A/K_S
175-192	59-64	0.273	0.796
178-195	60-65	0.277	0.808
181-198	61-66	0.240	0.700
184-201	62-67	0.345	1.006
187-204	63-68	0.427	1.245
190-207	64-69	0.169	0.493
193-210	65-70	0.016	0.047
196-213	66-71	0.016	0.047
199-216	67-72	0.022	0.064
202-219	68-73	0.017	0.050
205-222	69-74	0.014	0.041
208-225	70-75	0.008	0.023
211-228	71-76	0.006	0.017
214-231	72-77	0.005	0.015
217-234	73-78	0.001	0.003
220-237	74-79	0.001	0.003
223-240	75-80	0.000	0.000
226-243	76-81	0.043	0.125
229-246	77-82	0.039	0.114
232-249	78-83	0.040	0.117
235-252	79-84	0.040	0.117
238-255	80-85	0.040	0.117
241-258	81-86	0.040	0.117
244-261	82-87	0.039	0.114
247-264	83-88	0.004	0.012
250-267	84-89	0.004	0.012
253-270	85-90	0.008	0.023
256-273	86-91	0.011	0.032
259-276	87-92	0.131	0.382
262-279	88-93	0.081	0.236
265-282	89-94	0.015	0.044
268-285	90-95	0.023	0.067
271-288	91-96	0.086	0.251
274-291	92-97	0.022	0.064
277-294	93-98	0.019	0.055
280-297	94-99	0.016	0.047
283-300	95-100	0.011	0.032
286-303	96-101	0.004	0.012
289-306	97-102	0.003	0.009
292-309	98-103	0.004	0.012
295-312	99-104	0.003	0.009
298-315	100-105	0.003	0.009
301-318	101-106	0.044	0.128
304-321	102-107	0.041	0.120
307-324	103-108	0.009	0.026
310-327	104-109	0.009	0.026
313-330	105-110	0.043	0.125
316-333	106-111	0.047	0.137
319-336	107-112	0.100	0.292
322-339	108-113	0.015	0.044
325-342	109-114	0.016	0.047
328-345	110-115	0.016	0.047
331-348	111-116	0.005	0.015
334-351	112-117	0.018	0.052
337-354	113-118	0.202	0.589
340-357	114-119	0.038	0.111
343-360	115-120	0.021	0.061
346-363	116-121	0.085	0.248
349-366	117-122	0.051	0.149
352-369	118-123	0.031	0.090
355-372	119-124	0.031	0.090

Continued on next page

Table B.6 – continued from previous page

Window of the nucleotide positions	Corresponding amino acid positions	K_A	K_A/K_S
358-375	120-125	0.020	0.058
361-378	121-126	0.020	0.058
364-381	122-127	0.037	0.108
367-384	123-128	0.023	0.067
370-387	124-129	0.052	0.152
373-390	125-130	0.052	0.152
376-393	126-131	0.035	0.102
379-396	127-132	0.036	0.105
382-399	128-133	0.017	0.050
385-402	129-134	0.014	0.041
388-405	130-135	0.002	0.006
391-408	131-136	0.022	0.064
394-411	132-137	0.024	0.070
397-414	133-138	0.044	0.128
400-417	134-139	0.388	1.131
403-420	135-140	0.434	1.265
406-423	136-141	0.606	1.767
409-426	137-142	0.390	1.137
412-429	138-143	0.472	1.376
415-432	139-144	0.083	0.242
418-435	140-145	0.010	0.029
421-438	141-146	0.010	0.029
424-441	142-147	0.007	0.020
427-444	143-148	0.007	0.020
430-447	144-149	0.001	0.003
433-450	145-150	0.006	0.017
436-453	146-151	0.006	0.017
439-456	147-152	0.005	0.015
442-459	148-153	0.006	0.017
445-462	149-154	0.041	0.120
448-465	150-155	0.207	0.603
451-468	151-156	0.194	0.566
454-471	152-157	0.272	0.793
457-474	153-158	0.235	0.685
460-477	154-159	0.231	0.673
463-480	155-160	0.340	0.991
466-483	156-161	0.328	0.956
469-486	157-162	0.050	0.146
472-489	158-163	0.002	0.006
475-492	159-164	0.054	0.157
478-495	160-165	0.027	0.079
481-498	161-166	0.052	0.152
484-501	162-167	0.017	0.050
487-504	163-168	0.022	0.064
490-507	164-169	0.028	0.082
493-510	165-170	0.042	0.122
496-513	166-171	0.083	0.242
499-516	167-172	0.010	0.029
502-519	168-173	0.010	0.029
505-522	169-174	0.016	0.047
508-525	170-175	0.321	0.936
511-528	171-176	0.232	0.676
514-531	172-177	0.076	0.222
517-534	173-178	0.103	0.300
520-537	174-179	0.068	0.198
523-540	175-180	0.091	0.265
526-543	176-181	0.049	0.143
529-546	177-182	0.129	0.376
532-549	178-183	0.031	0.090
535-552	179-184	0.086	0.251
538-555	180-185	0.019	0.055

Continued on next page

Table B.6 – continued from previous page

Window of the nucleotide positions	Corresponding amino acid positions	K_A	K_A/K_S
541-558	181-186	0.049	0.143
544-561	182-187	0.043	0.125
547-564	183-188	0.078	0.227
550-567	184-189	0.072	0.210
553-570	185-190	0.072	0.210
556-573	186-191	0.071	0.207
559-576	187-192	0.031	0.090
562-579	188-193	0.031	0.090
565-582	189-194	0.018	0.052
568-585	190-195	0.031	0.090
571-588	191-196	0.004	0.012
574-591	192-197	0.004	0.012
577-594	193-198	0.021	0.061
580-597	194-199	0.003	0.009
583-600	195-200	0.044	0.128
586-603	196-201	0.042	0.122
589-606	197-202	0.060	0.175
592-609	198-203	0.132	0.385
595-612	199-204	0.052	0.152
598-615	200-205	0.128	0.373
601-618	201-206	0.087	0.254
604-621	202-207	0.093	0.271
607-624	203-208	0.077	0.224
610-627	204-209	0.162	0.472
613-630	205-210	0.044	0.128
616-633	206-211	0.021	0.061
619-636	207-212	0.055	0.160
622-639	208-213	0.056	0.163
625-642	209-214	0.113	0.329
628-645	210-215	0.133	0.388
631-648	211-216	0.028	0.082
634-651	212-217	0.028	0.082
637-654	213-218	0.002	0.006
640-657	214-219	0.001	0.003
643-660	215-220	0.001	0.003
646-663	216-221	0.002	0.006
649-666	217-222	0.003	0.009
652-669	218-223	0.076	0.222
655-672	219-224	0.020	0.058
658-675	220-225	0.043	0.125
661-678	221-226	0.038	0.111
664-681	222-227	0.192	0.560
667-684	223-228	0.360	1.050
670-687	224-229	0.272	0.793
673-690	225-230	0.089	0.259
676-693	226-231	0.139	0.405
679-696	227-232	0.144	0.420
682-699	228-233	0.210	0.612
685-702	229-234	0.275	0.802
688-705	230-235	0.001	0.003
691-708	231-236	0.005	0.015
694-711	232-237	0.011	0.032
697-714	233-238	0.028	0.082
700-717	234-239	0.034	0.099
703-720	235-240	0.036	0.105
706-723	236-241	0.086	0.251
709-726	237-242	0.116	0.338
712-729	238-243	0.064	0.187
715-732	239-244	0.004	0.012
718-735	240-245	0.012	0.035
721-738	241-246	0.163	0.475

Continued on next page

Table B.6 – continued from previous page

Window of the nucleotide positions	Corresponding amino acid positions	K_A	K_A/K_S
724-741	242-247	0.001	0.003
727-744	243-248	0.000	0.000
730-747	244-249	0.028	0.082
733-750	245-250	0.028	0.082
736-753	246-251	0.076	0.222
739-756	247-252	0.076	0.222
742-759	248-253	0.078	0.227
745-762	249-254	0.073	0.213
748-765	250-255	0.078	0.227
751-768	251-256	0.079	0.230
754-771	252-257	0.027	0.079
757-774	253-258	0.308	0.898
760-777	254-259	0.181	0.528
763-780	255-260	0.144	0.420
766-783	256-261	0.199	0.580
769-786	257-262	0.005	0.015
772-789	258-263	0.007	0.020
775-792	259-264	0.000	0.000
778-795	260-265	0.000	0.000
781-798	261-266	0.000	0.000
784-801	262-267	0.000	0.000
787-804	263-268	0.000	0.000

Table B.7: Energy values determined from the tertiary structures of the existing 190 fHbp sequences. Below are the explanations for the table:
Column 1: name of the sequence of which the tertiary structure was predicted and energy values determined.
Column 2: template used by SWISS-MODEL to predict the structure of the corresponding sequence in column 1. Here SM is the abbreviation for SWISS-MODEL.
Column 3: percentage identity of the sequence in column 1 with the template used in column 2.
Column 4: energy of the structure determined by SWISS-MODEL (in KJ/mol).
Column 5: energy value after energy minimization was performed on the structures by Swiss-PdbViewer (in KJ/mol). Here SPDBV is the abbreviation for Swiss-PdbViewer.
Column 6: energy determined by FoldX before energy minimization was performed on the structure (Kcal/mol).
Column 7: energy determined by FoldX after energy minimization was performed on the structure (Kcal/mol).
Sequence structures with the highest energy values (after minimization) are highlighted with yellow colour.

Sequence	**Template-SM**	**%identity**	**Energy-SM**	**Energy minimization-SPDBV**	**Energy-FoldX**	
					Before minimization	**After minimization**
B6EAW6	2W80C	94.191	-10531.661	-11133.833	131.55	121.181
C0JFM7	2W80C	95.436	-10616.156	-11152.968	129.99	113.61
B5U0Q5	2W80C	93.361	-10756.280	-11286.105	126.96	113.19
B2CQ01	2W80C	93.361	-10482.681	-11057.957	128.71	113.50
C0JFL9	2W80C	92.946	-10549.528	-11019.172	134.81	119.62
C0JF88	2W80C	92.946	-10577.336	-11155.581	132.48	117.25
B2CQ02	2W80C	93.776	-10762.610	-11350.365	129.43	113.21
Q6VRY3	2W80C	93.361	-10226.646	-10867.855	139.00	118.39
C0JFJ0	2W80C	92.946	-10224.869	-10800.083	128.44	113.77
B2CQ14	2W80C	92.531	-10217.853	-10793.306	129.08	113.08
C0JFI0	2W80C	92.116	-10777.321	-11315.262	134.24	118.43
C0JFJ1	2W80C	92.531	-10713.425	-11364.190	134.73	117.40
C0JF55	2W80C	92.531	-10739.799	-11378.031	134.85	118.06
C0JFF4	2W80C	92.946	-10800.684	-11454.406	136.43	119.37
C0JFJ5	2W80C	92.531	-10631.487	-11159.812	140.56	125.89
C0JFK9	2W80C	93.361	-10820.896	-11463.966	132.85	116.72
C0JF77	2W80C	93.361	-10656.380	-11304.994	136.01	118.00
C0JFJ8	2W80C	93.776	-10712.312	-11290.231	126.40	112.21
C0JFC9	2W80C	93.776	-11133.926	-11780.069	136.62	117.87
C0JF66	2W80C	94.606	-10941.952	-11591.748	134.21	118.14
C0JF56	2W80C	92.531	-10578.812	-11146.649	136.18	119.43
C0JFD7	2W80C	92.531	-10578.812	-11146.649	136.18	119.43
C0JF86	2W80C	91.701	-10215.619	-10892.890	138.41	120.61
C0JFN8	2W80C	91.701	-10652.901	-11232.785	135.33	118.74
C0JF51	2W80C	90.456	-10206.750	-10762.514	138.03	119.69
C0JFC8	2W80C	90.871	-10454.724	-10973.460	137.99	122.80
Q6QCB7	2W80C	91.286	-10318.200	-10995.616	141.92	123.67
B6EAW9	2W80C	92.531	-10095.026	-10768.841	136.98	117.69
C0JF69	2W80C	92.531	-10120.293	-10790.710	137.87	117.60
C0JFC7	2W80C	92.116	-10137.560	-10817.146	139.83	119.90
C0JFL7	2W80C	92.116	-10137.560	-10817.146	139.83	119.90
C0JFF1	2W80C	92.946	-10425.335	-11107.412	140.58	122.35
C0JFD0	2W80C	92.531	-10024.691	-10698.051	138.03	118.68
C0JFD1	2W80C	94.191	-10359.649	-10843.340	140.58	117.38
C0JFM1	2W80C	94.606	-10928.335	-11460.077	138.03	116.14
C0JF78	2W80C	94.606	-11774.460	-12261.739	132.19	116.51
C0JFA0	2W80C	94.606	-11876.771	-12363.976	131.14	120.33
C0JFN4	2W80H	92.116	-11272.427	-11725.354	130.61	124.14
B6EAW8	2W80C	92.531	-10891.298	-11408.136	134.45	120.12

Continued on next page

Table B.7 – continued from previous page

Sequence	Template-SM	%identity	Energy-SM	Energy minimization-SPDBV	Energy-FoldX	
					Before minimization	After minimization
Q6VRY2	2W80C	95.851	-10767.896	-11415.429	132.32	115.56
B5AAS3	2W80C	96.266	-10624.971	-11275.423	130.67	113.41
C0JFD9	2W80C	96.266	-10575.316	-11236.889	131.29	111.25
C0JF76	2W80C	96.680	-10623.077	-11275.744	129.70	111.45
B2CQ04	2W80C	94.606	-10613.559	-11253.764	139.36	122.85
C0JFD6	2W80C	96.680	-11072.190	-11559.334	129.59	115.15
Q6VRZ1	2W80C	97.095	-11032.109	-11512.100	127.31	114.22
C0JFF9	2W80C	96.680	-10814.681	-11293.024	125.75	112.41
C0JFH9	2W80C	97.925	-10699.144	-11341.228	126.80	110.79
C0JFN3	2W80C	96.680	-10879.147	-11522.341	127.00	111.77
C0JFG5	2W80H	95.436	-10752.379	-11294.184	128.34	118.61
C0JFJ6	2W80H	94.191	-10738.342	-11266.279	129.68	118.08
C0JF60	2W80C	97.095	-10645.220	-11294.862	131.24	115.47
Q6VRY1	2W80C	96.680	-10938.603	-11417.902	131.37	117.79
C0JFK6	2W80C	96.680	-11087.590	-11562.205	130.53	117.79
C0JF52	2W80C	96.680	-11002.455	-11483.229	128.11	114.69
C0JFH8	2W80C	95.021	-11437.172	-11947.255	128.33	115.08
C0JFC1	2W80C	96.680	-10639.744	-11116.229	124.88	111.72
C0JF53	2W80C	97.925	-10918.848	-11396.784	125.62	111.41
C0JFI8	2W80C	97.510	-11013.282	-11495.131	129.56	116.38
Q6VS09	2W80H	95.833	-11349.701	-11769.615	126.79	121.20
A1IQ30	2W80H	95.833	-11349.701	-11769.615	126.79	121.20
Q6QCB6	2W80C	95.851	-11338.722	-11817.962	133.33	120.42
C0JFJ3	2W80C	95.021	-11190.874	-11673.789	134.07	120.59
C0JF57	2W80C	95.436	-11690.869	-12171.490	133.52	120.51
C0JFN1	2W80H	95.021	-11238.752	-11655.619	126.43	121.98
C0JFH0	2W80C	95.851	-11583.762	-12065.236	133.27	119.89
C0JF82	2W80C	96.266	-11164.731	-11644.607	133.36	120.18
C0JFE6	2W80C	95.851	-11369.610	-11868.643	133.57	119.02
C0JFA4	2W80C	96.266	-11441.965	-11919.707	125.53	112.22
C0JF68	2W80H	100.000	-11180.135	-11612.103	123.74	115.46
C0JFL8	2W80C	99.170	-11426.974	-11883.679	123.66	112.17
Q6VRX9	2W80C	99.585	-11025.031	-11591.768	130.30	117.44
C0JF49	2W80C	99.585	-11216.915	-11691.964	127.49	114.78
Q6QCC2	2W80C	100.000	-11121.699	-11593.374	123.58	110.86
Q9JXV4	2W80C	100.000	-11121.699	-11621.195	125.16	117.99
C0JF85	2W80C	99.585	-11431.713	-11906.461	124.68	111.88
C0JF89	2W80C	99.585	-11200.402	-11669.392	123.66	110.02
C0JFG6	2W80C	99.170	-11393.284	-11912.731	124.38	111.94
C0JFM8	2W80C	99.170	-11033.789	-11514.570	126.84	112.90
C0JFA3	2W80H	97.925	-10832.818	-11274.697	127.31	119.86
C0JFI1	2W80C	98.340	-11142.247	-11617.897	126.80	113.79
C0JF54	2W80C	92.946	-11062.328	-11583.791	130.18	115.80
B2CQ11	2W80C	95.851	-10874.896	-11392.997	126.79	113.50
Q6QCB9	2W80C	94.191	-11140.438	-11789.780	135.45	118.13
Q6QCC0	2W80C	95.021	-10499.775	-11158.187	136.90	118.04
B2CQ06	2W80C	95.021	-10603.600	-11054.826	122.82	111.12
B2CQ05	2W80C	95.021	-10834.814	-11485.610	131.07	114.60
C0JF81	2W80C	83.402	-10499.725	-11052.418	145.54	129.67
C0JF61	2W80C	67.635	-10686.848	-11342.188	178.79	143.38
B6EAW7	2W80C	67.635	-10754.555	-11363.701	177.51	147.42
C0JFE5	2W80C	65.975	-10625.757	-11248.660	180.22	150.46
C0JF65	2W80C	66.805	-10958.096	-11610.725	185.92	151.06
A1KS29	2W80C	67.220	-10844.030	-11457.808	181.22	150.53
Q15I62	2W80C	67.220	-10844.030	-11457.808	181.22	150.53
C0JFB4	2W80C	67.635	-10806.710	-11450.919	182.26	147.84
C0JFN0	2W80C	67.220	-10798.854	-11415.437	179.14	148.10
B9VX97	2W80C	67.635	-10895.406	-11538.047	184.19	149.56
B9VX91	2W80C	65.975	-11020.517	-11752.468	183.57	151.90
C0JFF2	2W80C	65.560	-11151.741	-11778.542	187.04	154.39

Continued on next page

Table B.7 – continued from previous page

Sequence	Template-SM	%identity	Energy-SM	Energy minimization-SPDBV	Energy-FoldX	
					Before minimization	After minimization
C0JF74	2W80C	66.390	-11002.064	-11649.442	182.66	150.11
B9VX98	2W80C	66.390	-11019.174	-11668.050	180.99	146.08
C0JF75	2W80C	66.390	-10899.526	-11630.599	182.11	148.14
C0JF79	2W80C	66.390	-10565.960	-11187.553	180.32	148.69
C0JFD5	2W80C	67.220	-10479.586	-11098.334	181.01	152.28
C0JF59	2W80C	67.635	-10806.710	-11450.919	182.26	147.84
C0JFH4	2W80C	67.220	-10878.382	-11521.309	181.80	146.51
C0JFB3	2W80C	67.635	-10895.406	-11538.047	184.19	149.56
C0JFN6	2W80C	68.050	-10791.079	-11437.161	181.04	146.12
C0JF97	2W80C	67.635	-9337.645	-10777.181	220.77	176.65
C0JFB2	2W80C	68.050	-10551.379	-11196.111	183.88	149.51
B7U1A0	2W80C	70.833	-9711.880	-10450.872	180.82	149.25
Q19KF8	2W80C	70.954	-9750.944	-10492.463	182.45	151.22
C0JFF5	2W80C	70.539	-9675.944	-10394.493	178.78	149.78
B9VX96	2W80C	70.954	-9750.944	-10492.463	182.45	151.22
C0JFE7	2W80C	70.539	-9846.163	-10600.577	182.98	152.03
Q6VS14	2W80C	70.539	-9789.599	-10523.953	184.70	153.80
Q6VS12	2W80C	71.369	-9751.148	-10400.058	181.37	149.04
C0JF90	2W80C	70.124	-10033.418	-10662.467	181.79	160.35
Q6VS11	2W80C	70.954	-9831.283	-10581.907	182.38	151.14
C0JF87	2W80C	70.954	-9638.568	-10384.189	179.50	147.79
B9VX93	2W80C	70.539	-10040.818	-10782.177	182.29	150.82
C0JFA9	2W80C	70.124	-10017.021	-10651.178	184.69	154.28
C0JFA5	2W80C	70.124	-10036.527	-10756.798	182.01	152.49
B2CQ00	2W80C	70.539	-10040.818	-10782.177	182.29	150.82
C0JFC4	2W80C	70.539	-9959.590	-10697.271	178.37	149.26
B9VX82	2W80C	70.954	-10026.688	-10767.900	181.12	150.28
C0JFG8	2W80C	70.539	-9927.947	-10672.065	178.98	147.37
B5U0Q7	2W80C	70.124	-9859.055	-10484.930	175.18	145.59
C0JF96	2W80C	70.954	-9937.392	-10679.862	179.01	147.60
B5U0Q4	2W80C	91.701	-9910.618	-10423.753	141.02	124.76
C0JFH5	2W80C	90.871	-9877.449	-10480.127	140.45	122.71
C0JFI9	2W80C	91.286	-9906.229	-10424.311	141.30	123.97
C0JF99	2W80H	90.871	-10310.191	-10815.887	147.29	135.52
C0JF84	2W80C	91.701	-10761.366	-11329.370	136.95	121.02
A9M1G0	2W80C	96.266	-11654.548	-12173.951	131.72	118.38
C0JF58	2W80C	95.021	-10822.093	-11323.104	134.70	119.53
C0JFH2	2W80C	93.776	-11425.279	-11946.477	132.43	118.54
C0JF64	2W80C	87.552	-9866.758	-10457.859	149.48	131.81
Q6VRY4	2W80C	87.967	-10012.252	-10621.463	150.49	132.43
C0JFD3	2W80C	87.552	-9864.772	-10390.163	147.16	128.68
A1E5L5	2W80C	86.885	-9758.841	-10523.716	190.52	165.57
C0JFD4	2W80C	86.885	-9758.841	-10523.716	190.52	165.57
B9VX99	2W80C	87.295	-9851.169	-10613.350	190.47	165.69
C0JFF3	2W80H	88.115	-9932.995	-10512.156	173.97	158.45
B2CQ09	2W80C	88.525	-9833.204	-10563.511	190.54	163.91
B9VX94	2W80C	84.836	-9371.489	-10295.334	195.16	168.19
C0JF62	2W80C	86.885	-8871.511	-9973.303	189.04	162.56
B2CPZ9	2W80C	62.705	-8128.872	-9093.258	228.75	190.60
Q15I64	2W80C	62.295	-8132.525	-9097.169	229.84	191.74
B5LY76	2W80C	62.295	-8143.680	-9109.000	229.78	191.62
C0JF91	2W80C	61.885	-8142.168	-9337.731	221.55	176.71
C0JF83	2W80C	62.705	-8209.204	-9168.207	227.14	190.61
C0JFC6	2W80C	63.115	-8136.478	-9102.779	224.73	187.10
B9VX85	2W80C	62.705	-8128.872	-9093.258	228.75	190.60
B9VX90	2W80C	61.885	-8591.852	-9518.320	232.55	194.89
C0JFG2	2W80C	61.885	-8512.559	-9436.139	230.32	193.07
C0JF80	2W80C	62.295	-8646.759	-9560.436	226.42	191.31
C0JFM9	2W80C	62.295	-8646.759	-9560.436	226.42	191.31
C0JFI2	2W80C	62.295	-8650.425	-9563.077	226.17	190.84

Continued on next page

Table B.7 – continued from previous page

Sequence	Template-SM	%identity	Energy-SM	Energy minimization-SPDBV	Energy-FoldX	
					Before minimization	After minimization
C0JFA6	2W80C	62.705	-8494.505	-9460.716	226.22	189.27
B9VX89	2W80C	62.705	-8404.295	-9362.868	224.32	187.53
C0JFA8	2W80C	62.705	-8315.057	-9269.160	222.32	186.44
C0JF63	2W80C	62.705	-8473.467	-9437.166	228.97	191.02
C0JFI7	2W80C	63.115	-8480.316	-9445.769	225.09	188.61
C0JFG3	2W80C	61.885	-8720.467	-9673.294	232.87	198.41
B9VX84	2W80C	63.115	-8136.478	-9102.779	224.73	187.10
B9VX92	2W80C	62.705	-8507.856	-9464.061	224.06	187.49
C0JFB9	2W80C	62.705	-8256.821	-9214.522	225.07	188.23
C0JFK2	2W80C	63.115	-8352.843	-9315.410	225.11	188.81
C0JFL0	2W80C	61.885	-8888.626	-9823.227	231.49	194.52
C0JFE9	2W80C	62.295	-8664.627	-9582.113	227.58	192.02
B5U0Q6	2W80C	62.295	-8664.627	-9582.113	227.58	192.02
C0JFC3	2W80C	61.885	-8369.775	-9300.877	234.15	198.78
B5L521	2W80C	60.246	-9166.502	-10132.031	225.19	186.03
B2CQ08	2W80H	60.656	-7830.792	-9419.820	248.48	201.10
C0JFJ9	2W80H	59.426	-8010.897	-9623.688	264.59	212.15
C0JFL1	2W80H	59.836	-7645.381	-9351.549	255.54	209.85
C0JF92	2W80H	59.016	-7697.173	-9390.774	256.25	211.29
C0JFB0	2W80H	59.426	-7499.236	-9196.076	254.35	208.66
C0JFM6	2W80H	59.426	-7499.236	-9196.076	254.35	208.66
Q6VS28	2W80H	59.426	-7142.289	-8840.635	254.44	209.27
C0JFF0	2W80H	59.426	-7846.313	-9539.544	255.39	209.41
C0JFJ4	2W80H	59.426	-7434.237	-9145.395	257.86	211.38
C0JFM5	2W80H	59.016	-7786.667	-9509.900	259.92	212.76
B6VAX7	2W80H	59.426	-7846.313	-9539.544	255.39	209.41
Q19KF7	2W80H	58.607	-9007.075	-9956.911	222.02	189.29
C0JFL2	2W80H	59.836	-7518.411	-9213.378	256.73	210.61
C0JFP2	2W80H	59.016	-8843.925	-10302.643	244.47	201.27
C0JFM3	2W80H	59.836	-7600.051	-9256.514	253.81	210.34
C0JFM0	2W80C	64.344	-8766.095	-9704.785	229.93	193.77

Table B.8: Energy values determined from the tertiary structures of the 200 new variants of Set A. Refer back to the caption for Table B.7 for an explanation of the columns.

Sequence	**Template-SM**	**%identity**	**Energy-SM**	**Energy minimization-SPDBV**	**Energy-FoldX**		Region
					Before minimization	After minimization	
variant1	2W80H	80.083	-9457.933	-9945.355	155.51	145.51	Black
variant2	2W80C	72.951	-9148.821	-10139.492	203.99	172.83	Blue
variant3	2W80C	79.918	-9286.078	-10195.232	190.76	165.75	Blue
variant4	2W80C	70.539	-9765.071	-10472.156	179.61	150.63	Black
variant5	2W80C	92.946	-10782.896	-11328.200	137.16	121.07	Black
variant6	2W80C	83.402	-10795.215	-11311.810	157.47	139.60	Black
variant7	2W80C	88.797	-11271.610	-11867.308	141.69	122.14	Black
variant8	2W80C	75.104	-9990.903	-10505.535	170.14	149.45	Black
variant9	2W80C	80.328	-10143.482	-10820.600	198.71	170.82	Black
variant10	2W80C	91.701	-10578.847	-11209.582	139.21	122.80	Black
variant11	2W80C	91.393	-10125.021	-11027.775	185.58	163.96	Blue
variant12	2W80C	82.573	-10390.766	-10934.659	161.81	142.04	Black
variant13	2W80C	79.253	-10622.531	-11184.693	156.47	130.53	Black
variant14	2W80C	83.817	-11359.850	-11924.604	148.33	131.85	Black
variant15	2W80C	94.606	-10839.326	-11323.919	131.48	117.37	Black
variant16	2W80C	68.465	-9911.778	-10540.377	177.60	147.08	Black
variant17	2W80H	82.158	-10224.146	-10727.066	149.54	136.67	Black
variant18	2W80C	79.253	-10634.650	-11129.028	161.66	143.20	Black
variant19	2W80H	83.402	-10299.726	-10764.678	147.09	133.14	Black
variant20	2W80C	95.436	-10578.972	-11082.401	138.44	122.37	Black
variant21	2W80H	70.082	-9530.610	-10469.700	217.85	189.89	Red
variant22	2W80C	67.213	-10116.630	-11098.079	224.47	192.23	Red
variant23	2W80C	87.967	-10535.107	-11078.983	149.69	133.12	Black
variant24	2W80C	72.541	-9358.739	-10301.488	206.75	175.15	Blue
variant25	2W80C	79.253	-10472.126	-11063.370	160.83	134.74	Black
variant26	2W80C	80.498	-10210.012	-10760.080	162.10	143.39	Black
variant27	2W80C	84.232	-10376.513	-10981.525	147.41	125.08	Black
variant28	2W80C	80.738	-10047.333	-10705.186	192.61	166.95	Blue
variant29	2W80C	80.738	-9668.347	-10439.609	197.77	169.25	Blue
variant30	2W80H	85.246	-8725.480	-9519.747	176.03	160.69	Blue
variant31	2W80C	81.743	-10050.083	-10589.047	151.43	125.66	Black
variant32	2W80C	80.913	-10139.267	-10674.039	161.82	143.68	Black
variant33	2W80C	84.426	-9201.427	-10122.076	197.89	174.19	Blue
variant34	2W80C	76.349	-9939.480	-10474.274	162.51	137.77	Black
variant35	2W80C	90.456	-11309.157	-11810.707	139.23	120.26	Black
variant36	2W80C	83.402	-11063.755	-11560.197	156.69	136.95	Black
variant37	2W80C	82.988	-10452.530	-10963.493	146.08	127.21	Black
variant38	2W80H	79.253	-9560.660	-10008.704	153.33	144.50	Black
variant39	2W80C	72.614	-9877.507	-10506.573	177.75	148.16	Black
variant40	2W80C	79.098	-9784.690	-10537.945	194.06	165.89	Blue
variant41	2W80C	69.262	-9843.384	-10657.755	210.00	176.06	Blue
variant42	2W80C	87.705	-9622.698	-10284.426	182.57	160.07	Blue
variant43	2W80C	80.913	-10015.891	-10607.223	158.50	132.32	Black
variant44	2W80C	80.083	-10254.484	-10755.518	150.67	133.16	Black
variant45	2W80C	89.212	-10802.322	-11387.605	145.49	126.78	Black
variant46	2W80C	79.508	-9466.285	-10132.309	200.30	168.56	Blue
variant47	2W80C	66.803	-10060.303	-11039.364	219.05	186.27	Red
variant48	2W80C	88.525	-9588.789	-10382.045	190.19	163.10	Blue
variant49	2W80C	83.817	-10855.912	-11359.273	157.66	138.46	Black
variant50	2W80C	94.191	-10846.040	-11373.340	131.68	116.91	Black
variant51	2W80C	69.710	-10394.806	-11021.561	174.92	142.75	Black
variant52	2W80H	66.393	-8531.624	-9475.286	225.46	195.52	Red
variant53	2W80C	82.988	-9818.149	-10337.633	155.28	131.29	Black
variant54	2W80C	77.593	-9921.555	-10721.499	165.95	137.91	Black
variant55	2W80C	68.050	-9368.602	-10077.069	179.44	146.92	Black
variant56	2W80C	65.145	-9079.731	-9768.427	176.93	146.66	Black
variant57	2W80C	93.361	-10486.833	-11144.964	132.72	116.08	Black

Continued on next page

Table B.8 – continued from previous page

Sequence	Template-SM	%identity	Energy-SM	Energy minimization-SPDBV	Energy-FoldX		Region
					Before minimization	After minimization	
variant58	2W80C	83.607	-9536.220	-10385.263	200.56	174.80	Blue
variant59	2W80C	80.498	-10341.617	-10892.579	161.24	139.73	Black
variant60	2W80C	84.583	-10734.181	-11277.963	146.97	131.15	Black
variant61	2W80C	70.539	-9459.084	-10189.881	173.58	142.47	Black
variant62	2W80C	80.833	-10124.169	-10658.074	162.92	145.48	Black
variant63	2W80C	82.787	-11777.827	-12379.118	193.57	166.08	Blue
variant64	2W80C	78.689	-9902.706	-10573.473	201.68	173.53	Blue
variant65	2W80C	67.220	-9790.835	-10542.499	187.84	152.11	Black
variant66	2W80C	92.531	-11396.927	-11876.922	131.18	117.47	Black
variant67	2W80C	78.279	-10223.786	-11021.528	186.33	157.44	Black
variant68	2W80C	87.967	-11550.169	-12123.965	150.75	131.09	Black
variant69	2W80C	92.531	-11501.028	-12163.365	137.91	121.19	Black
variant70	2W80C	77.178	-10886.383	-11380.337	162.31	142.22	Black
variant71	2W80C	85.062	-11578.486	-12087.947	149.08	131.89	Black
variant72	2W80C	68.465	-9352.567	-9963.908	177.94	148.53	Black
variant73	2W80C	78.008	-11142.536	-11822.432	156.09	126.90	Black
variant74	2W80C	79.918	-10024.214	-10699.768	198.79	171.03	Blue
variant75	2W80C	70.082	-9237.715	-10084.507	212.45	177.56	Blue
variant76	2W80H	94.672	-10586.634	-11152.307	165.55	152.41	Black
variant77	2W80C	82.573	-9865.973	-10414.266	146.48	120.20	Black
variant78	2W80C	88.382	-10953.373	-11479.360	140.59	124.66	Black
variant79	2W80C	71.784	-9525.796	-10137.747	176.75	146.59	Black
variant80	2W80C	90.984	-9631.081	-10722.171	176.25	154.80	Black
variant81	2W80C	78.279	-9675.076	-10703.731	212.15	180.19	Red
variant82	2W80C	88.934	-10179.438	-11068.558	175.46	152.25	Black
variant83	2W80C	81.148	-9992.201	-10753.974	188.53	161.12	Blue
variant84	2W80C	75.519	-9367.917	-10080.765	167.62	140.74	Black
variant85	2W80C	71.369	-10160.604	-10803.292	177.61	147.65	Black
variant86	2W80C	90.456	-11369.580	-11875.500	133.40	116.97	Black
variant87	2W80C	78.008	-9564.075	-10072.911	159.35	135.26	Black
variant88	2W80C	82.917	-10580.610	-11115.575	150.71	133.09	Black
variant89	2W80C	89.627	-10316.643	-10816.772	143.78	127.43	Black
variant90	2W80H	86.667	-10703.701	-11153.046	137.43	128.73	Black
variant91	2W80C	92.531	-10220.166	-10868.675	132.10	115.95	Black
variant92	2W80C	96.266	-11163.879	-11662.473	132.76	119.62	Black
variant93	2W80C	89.627	-10012.005	-10542.592	148.11	130.92	Black
variant94	2W80C	93.443	-9869.062	-10692.230	167.08	150.83	Black
variant95	2W80H	89.627	-10116.369	-10574.310	144.83	137.26	Black
variant96	2W80C	84.232	-10337.072	-10850.841	155.12	135.41	Black
variant97	2W80C	79.668	-9833.333	-10518.075	156.57	131.02	Black
variant98	2W80C	90.456	-11585.327	-12006.833	125.89	113.62	Black
variant99	2W80C	79.253	-9672.446	-10242.239	164.11	137.64	Black
variant100	2W80H	77.178	-8657.629	-9984.699	189.45	156.50	Black
variant101	2W80C	80.328	-9442.451	-10110.702	200.01	169.77	Blue
variant102	2W80C	80.913	-10157.588	-10677.139	157.35	138.39	Black
variant103	2W80H	93.776	-10456.299	-10891.264	135.91	125.06	Black
variant104	2W80C	92.116	-11024.611	-11682.397	141.33	121.32	Black
variant105	2W80H	80.083	-9704.047	-10101.760	151.32	141.63	Black
variant106	2W80C	90.984	-9698.368	-10795.982	175.80	154.69	Black
variant107	2W80C	79.668	-11299.051	-11830.786	161.78	138.18	Black
variant108	2W80H	93.361	-10967.451	-11491.380	127.10	118.31	Black
variant109	2W80C	81.328	-10398.499	-10969.610	157.52	130.94	Black
variant110	2W80C	80.498	-11144.553	-11731.245	158.72	129.01	Black
variant111	2W80C	69.262	-9774.072	-10596.871	211.65	177.82	Blue
variant112	2W80C	95.021	-10736.114	-11258.754	131.70	119.17	Black
variant113	2W80C	83.817	-10454.758	-10970.132	146.22	129.27	Black
variant114	2W80H	89.627	-11579.053	-12090.739	140.31	122.82	Black
variant115	2W80C	74.180	-9298.311	-10070.895	179.59	160.74	Blue
variant116	2W80C	95.851	-10926.499	-11495.369	128.76	113.69	Black
variant117	2W80C	82.573	-10812.596	-11344.157	156.04	139.55	Black

Continued on next page

Table B.8 – continued from previous page

Sequence	Template-SM	%identity	Energy-SM	Energy minimization-SPDBV	Energy-FoldX		Region
					Before minimization	After minimization	
variant118	2W80C	70.539	-9776.823	-10409.971	179.05	155.01	Black
variant119	2W80C	76.763	-11224.499	-11686.066	174.58	150.65	Black
variant120	2W80C	69.262	-9941.857	-10742.342	216.66	186.15	Red
variant121	2W80C	82.573	-10168.320	-10800.557	150.30	125.79	Black
variant122	2W80C	78.279	-9590.757	-10461.796	193.70	164.92	Blue
variant123	2W80C	68.050	-10322.688	-11060.227	172.86	138.83	Black
variant124	2W80C	76.230	-9811.961	-10595.711	213.87	180.79	Red
variant125	2W80C	74.590	-9669.887	-10432.032	216.27	183.80	Red
variant126	2W80C	85.062	-10889.364	-11426.450	151.75	131.67	Black
variant127	2W80C	74.689	-9254.352	-9835.760	164.50	134.66	Black
variant128	2W80C	82.988	-10797.580	-11352.781	148.29	130.02	Black
variant129	2W80C	69.295	-10327.256	-10937.885	182.12	152.40	Black
variant130	2W80C	77.593	-9709.159	-10269.781	156.32	133.98	Black
variant131	2W80C	80.498	-9729.384	-10308.627	159.03	130.61	Black
variant132	2W80C	90.574	-10331.854	-10812.516	171.57	148.83	Black
variant133	2W80C	82.988	-11386.021	-11911.022	150.90	134.32	Black
variant134	2W80C	90.984	-10484.272	-11114.430	171.80	149.00	Black
variant135	2W80C	67.623	-10407.936	-11302.178	214.00	177.23	Blue
variant136	2W80H	91.286	-10400.424	-10844.545	134.30	122.75	Black
variant137	2W80C	94.606	-10793.630	-11318.522	130.66	116.56	Black
variant138	2W80C	87.137	-10997.760	-11571.583	146.31	127.96	Black
variant139	2W80H	80.913	-9722.173	-10153.371	150.48	141.83	Black
variant140	2W80C	76.349	-9355.457	-10017.917	167.31	141.16	Black
variant141	2W80C	92.946	-10733.037	-11262.052	139.15	120.80	Black
variant142	2W80C	65.560	-9114.008	-9740.824	184.44	148.97	Black
variant143	2W80C	80.328	-10222.184	-10766.485	166.94	144.54	Black
variant144	2W80C	70.539	-9709.394	-10430.348	179.44	146.09	Black
variant145	2W80C	89.627	-10655.434	-11204.207	133.10	118.08	Black
variant146	2W80C	85.000	-10642.542	-11157.885	142.35	126.91	Black
variant147	2W80C	89.627	-10126.970	-10748.699	140.20	121.96	Black
variant148	2W80C	73.029	-9599.265	-10235.366	175.34	146.74	Black
variant149	2W80C	71.784	-9816.910	-10540.073	178.70	151.40	Black
variant150	2W80C	89.754	-10471.438	-11064.831	180.00	158.67	Blue
variant151	2W80C	83.817	-10138.154	-10684.090	156.67	137.67	Black
variant152	2W80C	79.668	-9475.329	-9998.044	157.71	132.69	Black
variant153	2W80C	89.300	-10147.582	-10764.452	169.39	149.48	Black
variant154	2W80C	91.286	-11869.120	-12381.519	131.14	115.24	Black
variant155	2W80C	82.573	-10541.580	-11112.409	151.91	127.82	Black
variant156	2W80C	82.573	-10222.317	-10757.182	160.70	138.89	Black
variant157	2W80C	84.647	-10969.767	-11509.052	154.05	135.33	Black
variant158	2W80H	94.191	-10506.514	-11050.241	141.37	131.32	Black
variant159	2W80C	85.062	-11123.965	-11665.229	155.80	137.02	Black
variant160	2W80C	79.253	-10143.925	-10634.547	162.34	142.88	Black
variant161	2W80H	74.180	-9439.729	-10575.953	208.03	184.72	Red
variant162	2W80C	80.498	-10865.378	-11482.168	176.42	152.19	Black
variant163	2W80C	82.158	-9633.353	-10214.046	150.01	127.29	Black
variant164	2W80C	67.220	-8878.178	-9465.873	180.42	148.60	Black
variant165	2W80C	64.754	-8292.276	-9210.207	225.47	190.05	Red
variant166	2W80C	83.817	-11163.284	-11714.295	153.26	134.14	Black
variant167	2W80C	93.033	-10041.437	-10561.724	169.46	150.66	Black
variant168	2W80C	77.178	-9356.653	-9878.675	157.47	133.48	Black
variant169	2W80C	81.743	-9649.434	-10233.217	147.26	121.93	Black
variant170	2W80C	80.913	-9895.980	-10610.665	152.42	125.89	Black
variant171	2W80C	86.475	-9557.582	-10079.159	188.80	165.31	Blue
variant172	2W80C	87.967	-11843.673	-12335.377	138.63	120.91	Black
variant173	2W80C	81.743	-10701.622	-11248.795	163.80	144.28	Black
variant174	2W80C	89.212	-11220.054	-11777.958	132.10	115.41	Black
variant175	2W80C	95.021	-10590.702	-11129.076	131.18	117.79	Black
variant176	2W80C	78.008	-10564.250	-11117.474	174.26	154.68	Black
variant177	2W80C	79.098	-10790.891	-11471.730	201.26	171.23	Blue

Continued on next page

Table B.8 – continued from previous page

Sequence	**Template-SM**	**%identity**	**Energy-SM**	**Energy minimization-SPDBV**	**Energy-FoldX**		Region
					Before minimization	After minimization	
variant178	2W80C	75.104	-9448.995	-10046.504	160.05	134.30	Black
variant179	2W80C	92.116	-11301.126	-11856.587	133.39	114.84	Black
variant180	2W80C	92.946	-10847.402	-11332.817	126.75	112.91	Black
variant181	2W80C	81.743	-9924.828	-10454.911	163.24	140.54	Black
variant182	2W80C	66.393	-8616.766	-9450.605	223.03	187.05	Red
variant183	2W80C	93.361	-10754.855	-11412.120	135.69	119.06	Black
variant184	2W80C	91.701	-11922.575	-12350.325	130.29	119.38	Black
variant185	2W80C	74.180	-10634.179	-11416.469	212.24	179.72	Blue
variant186	2W80C	83.817	-10939.318	-11509.831	152.19	135.78	Black
variant187	2W80C	83.817	-11450.323	-12006.175	154.46	135.85	Black
variant188	2W80C	83.402	-10412.571	-10961.562	156.21	136.93	Black
variant189	2W80H	90.871	-10290.048	-10779.827	134.63	129.69	Black
variant190	2W80C	79.253	-9773.637	-10335.595	167.48	144.62	Black
variant191	2W80C	74.590	-10403.420	-11305.454	205.25	175.84	Blue
variant192	2W80C	92.531	-12087.654	-12651.916	135.13	118.85	Black
variant193	2W80C	89.212	-10422.977	-11068.375	151.52	133.28	Black
variant194	2W80C	90.871	-10175.699	-10636.776	141.22	125.96	Black
variant195	2W80C	84.647	-11002.570	-11515.552	154.23	134.72	Black
variant196	2W80H	75.934	-8430.079	-9862.659	189.40	156.14	Black
variant197	2W80C	87.552	-10159.515	-10662.250	147.47	134.90	Black
variant198	2W80C	83.817	-10776.937	-11291.300	143.68	129.44	Black
variant199	2W80C	84.647	-10089.458	-10686.725	154.53	133.90	Black
variant200	2W80H	77.178	-8602.641	-10021.542	194.34	159.87	Black

Table B.9: Energy values determined from the tertiary structures of the 100 new variants of Set B. Refer back to the caption for Table B.7 for an explanation of the columns.

Sequence	**Template-SM**	**%identity**	**Energy-SM**	**Energy minimization-SPDBV**	**Energy-FoldX**		Region
					Before minimization	After minimization	
variant1	2W80H	82.573	-10324.950	-10701.946	157.19	135.70	Black
variant2	2W80C	90.871	-10766.152	-11247.407	128.99	113.66	Black
variant3	2W80C	88.382	-10287.660	-10801.854	142.36	121.88	Black
variant4	2W80C	66.803	-9144.697	-10122.778	224.27	189.18	Red
variant5	2W80C	84.232	-10784.713	-11253.972	142.69	124.68	Black
variant6	2W80H	76.230	-8688.720	-9885.193	215.93	190.24	Red
variant7	2W80H	82.988	-8780.003	-10072.824	169.98	140.38	Black
variant8	2W80C	86.066	-9153.137	-10041.002	181.54	157.07	Blue
variant9	2W80C	75.104	-10081.688	-10800.198	170.61	139.97	Black
variant10	2W80C	87.705	-9890.827	-10854.473	182.40	156.30	Blue
variant11	2W80C	88.797	-10180.986	-10731.847	135.92	118.63	Black
variant12	2W80C	67.213	-10520.815	-11155.994	206.64	175.58	Blue
variant13	2W80C	92.531	-10672.493	-11182.649	134.02	118.82	Black
variant14	2W80C	78.838	-10809.700	-11329.611	169.44	145.84	Black
variant15	2W80C	82.573	-10059.699	-10723.966	156.85	136.23	Black
variant16	2W80C	80.498	-9631.497	-10221.550	158.62	132.99	Black
variant17	2W80C	82.158	-10295.935	-10774.081	157.24	142.09	Black
variant18	2W80C	67.635	-10697.323	-11231.710	175.49	148.88	Black
variant19	2W80C	88.525	-9369.943	-10531.088	183.40	159.19	Blue
variant20	2W80C	92.531	-9900.839	-10399.626	140.18	123.99	Black
variant21	2W80C	86.722	-11590.951	-11993.192	153.16	138.12	Black
variant22	2W80C	79.508	-10909.305	-11649.275	199.03	169.91	Blue
variant23	2W80C	89.212	-10446.812	-10930.201	137.48	120.15	Black
variant24	2W80C	89.212	-10105.470	-10737.102	141.07	124.83	Black
variant25	2W80C	86.885	-9950.098	-10764.620	184.60	161.09	Blue
variant26	2W80C	79.253	-10039.235	-10829.078	159.85	135.22	Black
variant27	2W80C	85.246	-9948.574	-10599.855	189.49	158.29	Blue
variant28	2W80H	82.158	-9009.153	-10304.185	175.31	146.16	Black
variant29	2W80C	68.050	-9646.900	-10267.302	179.39	148.59	Black
variant30	2W80H	76.763	-8498.539	-9834.258	191.72	158.63	Blue
variant31	2W80C	92.116	-10928.396	-11443.401	135.02	118.73	Black
variant32	2W80C	78.279	-10117.967	-10805.604	201.17	172.72	Blue
variant33	2W80C	89.627	-10847.556	-11316.542	145.92	132.35	Black
variant34	2W80C	80.738	-9342.032	-9958.600	185.27	159.67	Blue
variant35	2W80C	75.934	-10040.645	-10537.941	154.26	136.65	Black
variant36	2W80C	82.988	-11354.896	-11911.835	157.97	138.07	Black
variant37	2W80C	85.656	-9591.529	-10427.470	192.02	170.23	Blue
variant38	2W80C	83.607	-8069.898	-9268.037	203.79	172.65	Blue
variant39	2W80C	66.393	-9654.659	-10584.923	239.74	202.62	Red
variant40	2W80C	89.212	-10447.781	-11007.547	144.87	129.87	Black
variant41	2W80C	84.232	-10122.150	-10650.969	151.84	131.13	Black
variant42	2W80H	84.426	-8803.450	-9691.706	186.08	161.84	Blue
variant43	2W80C	82.158	-10244.059	-10819.255	169.40	141.09	Black
variant44	2W80C	83.197	-9767.897	-10523.331	186.12	161.56	Blue
variant45	2W80C	80.498	-10131.054	-10839.701	166.20	141.33	Black
variant46	2W80C	75.934	-9298.632	-9864.213	173.11	147.09	Black
variant47	2W80C	69.262	-10092.500	-11069.918	222.55	183.71	Red
variant48	2W80C	81.743	-11228.696	-11778.420	156.66	139.90	Black
variant49	2W80C	76.349	-10082.690	-10494.047	183.75	159.46	Blue
variant50	2W80C	91.701	-9990.165	-10519.723	129.85	113.26	Black
variant51	2W80H	72.951	-9717.099	-10820.308	209.05	182.31	Red
variant52	2W80C	75.820	-10430.614	-11338.894	204.02	175.59	Blue
variant53	2W80C	82.787	-9170.633	-9868.550	185.41	159.93	Blue
variant54	2W80C	76.639	-10689.628	-11461.009	213.04	178.64	Red
variant55	2W80C	84.232	-10783.727	-11311.987	148.17	130.46	Black
variant56	2W80C	83.402	-10733.672	-11295.898	154.91	136.65	Black
variant57	2W80C	90.041	-10821.503	-11366.747	139.69	124.61	Black

Continued on next page

Table B.9 – continued from previous page

Sequence	Template-SM	%identity	Energy-SM	Energy minimization-SPDBV	Energy-FoldX		Region
					Before minimization	After minimization	
variant58	2W80C	73.029	-9657.539	-10271.945	170.01	141.37	Black
variant59	2W80H	82.917	-9133.718	-10463.365	177.25	147.42	Black
variant60	2W80C	83.817	-11139.703	-11660.049	152.33	128.51	Black
variant61	2W80C	90.456	-10231.006	-10761.467	136.39	119.12	Black
variant62	2W80C	83.402	-10803.501	-11391.791	163.30	136.74	Black
variant63	2W80C	77.178	-10762.551	-11306.629	170.13	143.28	Black
variant64	2W80C	77.459	-9393.287	-10044.455	188.01	162.43	Blue
variant65	2W80C	87.552	-9992.338	-10516.071	147.59	130.39	Black
variant66	2W80C	83.402	-11780.542	-12312.651	151.86	131.77	Black
variant67	2W80C	90.871	-10781.665	-11304.812	144.45	128.53	Black
variant68	2W80H	81.743	-10691.781	-11143.228	163.37	143.72	Black
variant69	2W80C	81.148	-9589.752	-10396.552	208.57	179.90	Red
variant70	2W80C	75.934	-10519.759	-11109.483	169.59	142.30	Black
variant71	2W80C	75.410	-9310.462	-9987.796	207.87	175.00	Blue
variant72	2W80C	81.967	-9851.590	-10509.594	182.80	156.51	Blue
variant73	2W80C	91.701	-10260.119	-10801.245	126.29	115.16	Black
variant74	2W80C	81.328	-10382.327	-11087.752	159.87	137.77	Black
variant75	2W80C	68.050	-11192.534	-11911.011	180.68	148.05	Black
variant76	2W80C	68.443	-9807.025	-10699.413	213.31	179.54	Red
variant77	2W80H	69.295	-8715.495	-10101.164	214.29	174.84	Blue
variant78	2W80C	83.402	-10538.289	-11021.200	148.55	129.32	Black
variant79	2W80C	80.498	-10905.543	-11585.215	149.91	124.64	Black
variant80	2W80C	73.029	-9829.572	-10484.208	180.56	148.41	Black
variant81	2W80C	83.817	-9419.487	-9890.812	139.99	122.22	Black
variant82	2W80C	76.639	-9942.893	-10869.638	195.67	166.91	Blue
variant83	2W80C	75.820	-9720.432	-10369.436	207.24	175.51	Blue
variant84	2W80C	78.689	-10103.282	-10700.568	187.84	159.63	Blue
variant85	2W80C	82.573	-9866.408	-10385.570	153.95	138.59	Black
variant86	2W80C	79.668	-10514.534	-10967.464	147.80	129.82	Black
variant87	2W80C	90.041	-11060.084	-11576.387	142.84	128.39	Black
variant88	2W80C	79.098	-10149.094	-10898.811	201.05	173.57	Blue
variant89	2W80C	76.349	-10178.084	-10627.459	161.58	136.69	Black
variant90	2W80C	87.137	-9966.197	-10633.252	153.31	130.29	Black
variant91	2W80C	80.913	-10492.107	-11128.555	159.99	136.92	Black
variant92	2W80C	88.115	-10067.083	-10887.166	178.08	153.12	Black
variant93	2W80C	81.867	-10072.097	-10660.872	193.07	164.68	Blue
variant94	2W80H	79.253	-10090.178	-10544.397	157.72	145.33	Black
variant95	2W80C	80.498	-11032.714	-11654.353	166.09	137.35	Black
variant96	2W80C	82.787	-8672.978	-9618.117	197.15	172.25	Blue
variant97	2W80C	80.498	-11630.682	-12169.227	160.10	138.22	Black
variant98	2W80C	84.232	-11455.903	-11847.486	154.34	136.25	Black
variant99	2W80H	81.328	-9737.430	-10164.905	151.73	145.01	Black
variant100	2W80C	85.477	-10918.499	-11490.443	143.46	121.36	Black

Appendix C

Figures: Multiple Sequence Alignments

Figures C.1 to C.6 illustrate portions of the multiple sequence alignment of the 190 fHbp sequences including the signal protein portion. Figures C.7 to C.12 illustrate the various regions of the new variants of Set A. These figures lack the signal protein parts.

The signal protein parts (in Figures C.1 to C.6) and the various regions of the mature protein have been highlighted with different colours. The colour highlighting is co-ordinated across both sets of figures (for the 190 fHbp sequences and the new variants of Set A).

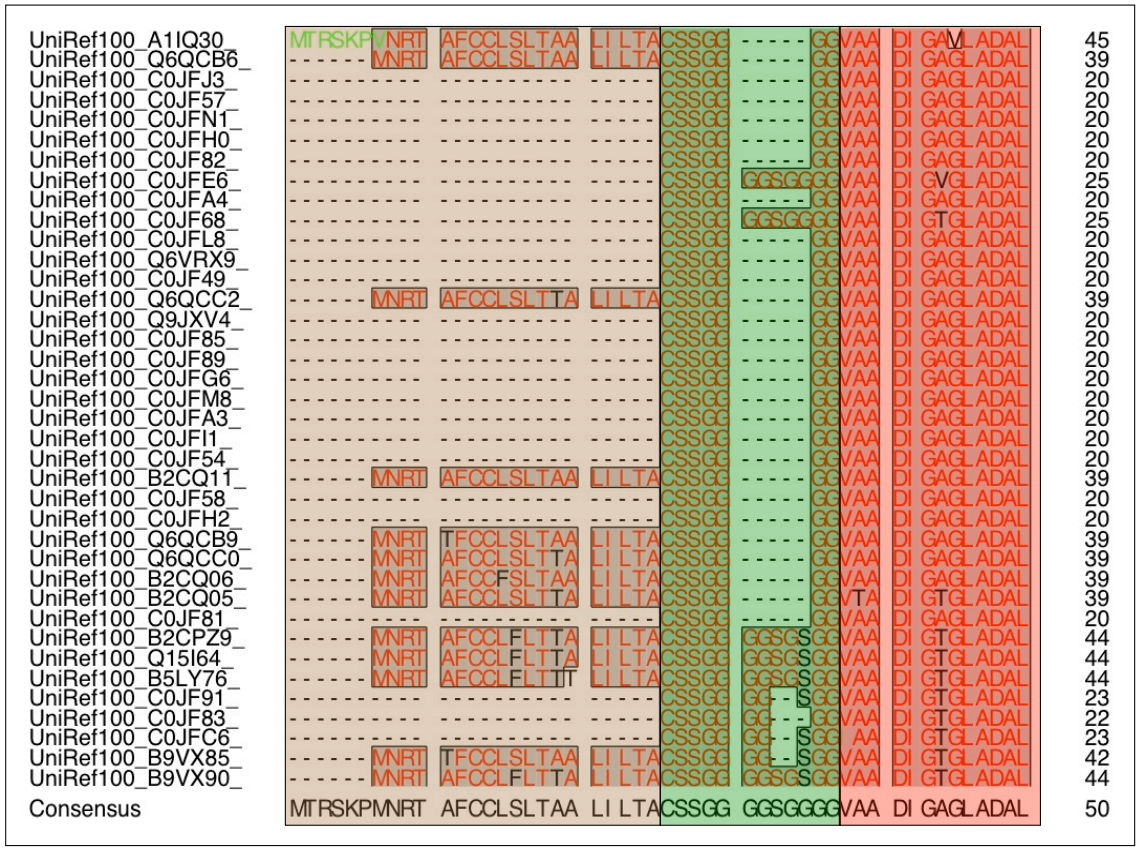


Figure C.1: A portion of the multiple sequence alignment of the 190 fHbp sequences illustrating the signal peptides (highlighted in brown colour), the amino-terminal repetitive region (highlighted in green colour) marking the beginning of the mature protein part (the N-term region of Figure 2.2), and the beginning of the first variable region V_A (highlighted in red) (Section 2.6 and Figure 2.2). The numbers on the right are the count of amino acids for individual sequence starting from the first amino acid of each sequence respectively; that is, the count starts from the first amino acid of the signal protein portion if it is present in a sequence or the first amino acid of the amino-terminal repetitive region, which will be a 'C'. The number excludes the count of 'gaps'.
The consensus sequence illustrated in this figure is obtained by using the `consensus` command of the EMBOSS *prettyplot* program. The command uses the sequence weights and a scoring matrix to calculate a score for each amino acid at each position in the alignment [4, 5]. The highest scoring residue goes into the consensus sequence if the score is higher than a user-specified `plurality` value, which in this case was set to 0. Hence, for positions where there was only one amino acid with the rest being 'gaps', that single amino acid appeared in the consensus sequence. Note that the consensus sequence illustrated here is derived from the alignment of the 190 fHbp sequences and the occurrences of 'gap' symbols are excluded while deriving the consensus sequence.

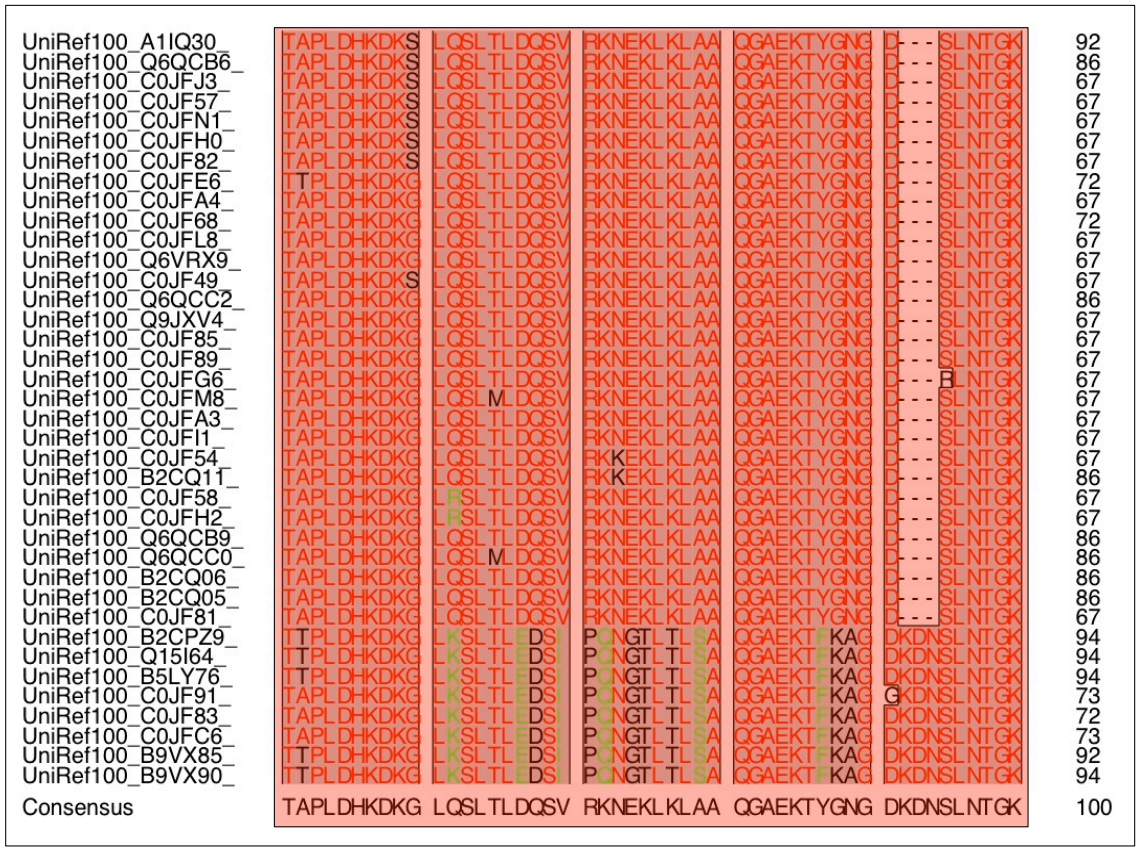


Figure C.2: A portion of the multiple sequence alignment of the 190 fHbp sequences illustrating the continuation of the variable region V_A. The numbers on the right again represent the count of the amino acids starting from the first amino acid in each individual sequence respectively. For instance, for the first sequence A1IQ30, the number on the right shows 92. This 92 is the sum of the amino acids that have occurred starting from the first amino acid of the signal part (amino acid 'M') in this sequence (c.f. Figure C.1) until this segment of the alignment.

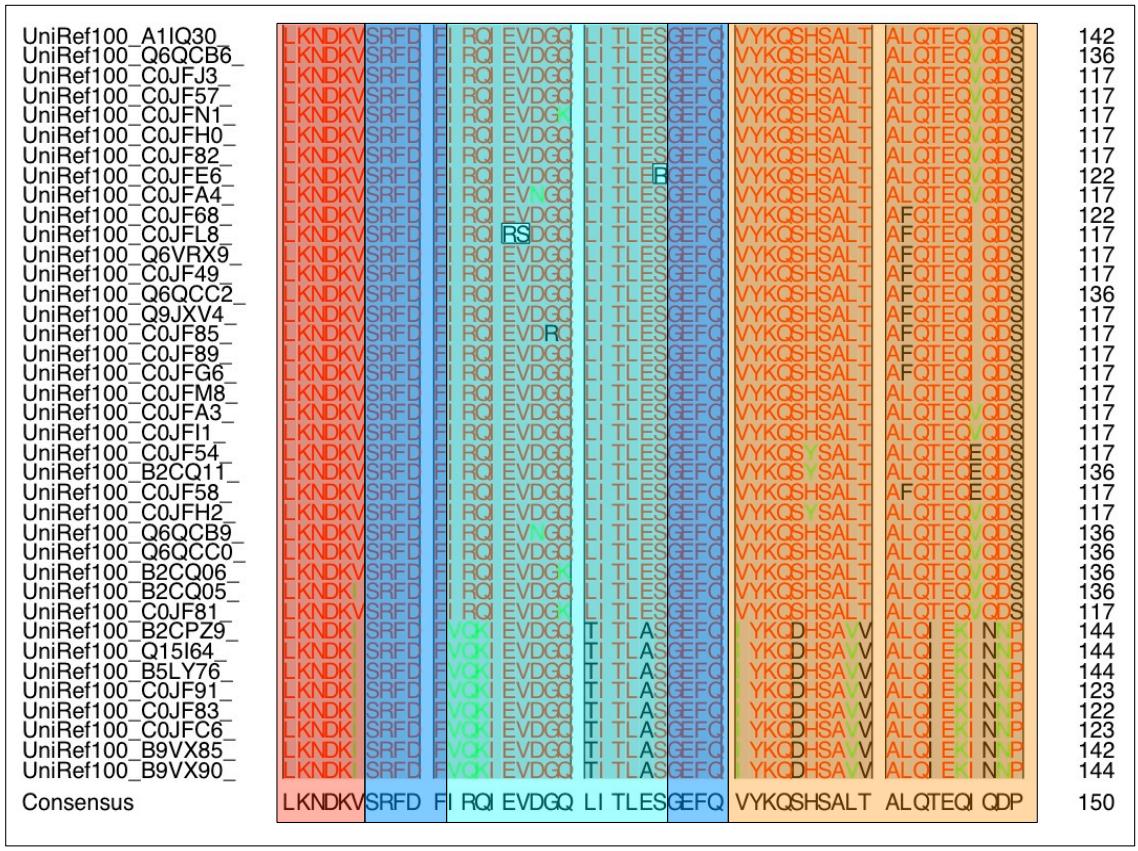


Figure C.3: A portion of the multiple sequence alignment of the 190 fHbp sequences illustrating the variable regions V_A, V_B, and V_C flanked by the invariant segments. The blue regions represent the invariant blocks and the other regions the modular variable ones. The red region, from the left, corresponds to the continuation of region V_A, region V_B is coloured in cyan, and V_C in orange. The blue regions ('SRFDF' and 'GEFQ') are the black vertical rectangles that separate the regions V_A - V_B and V_B - V_C, respectively, in Figure 2.2.

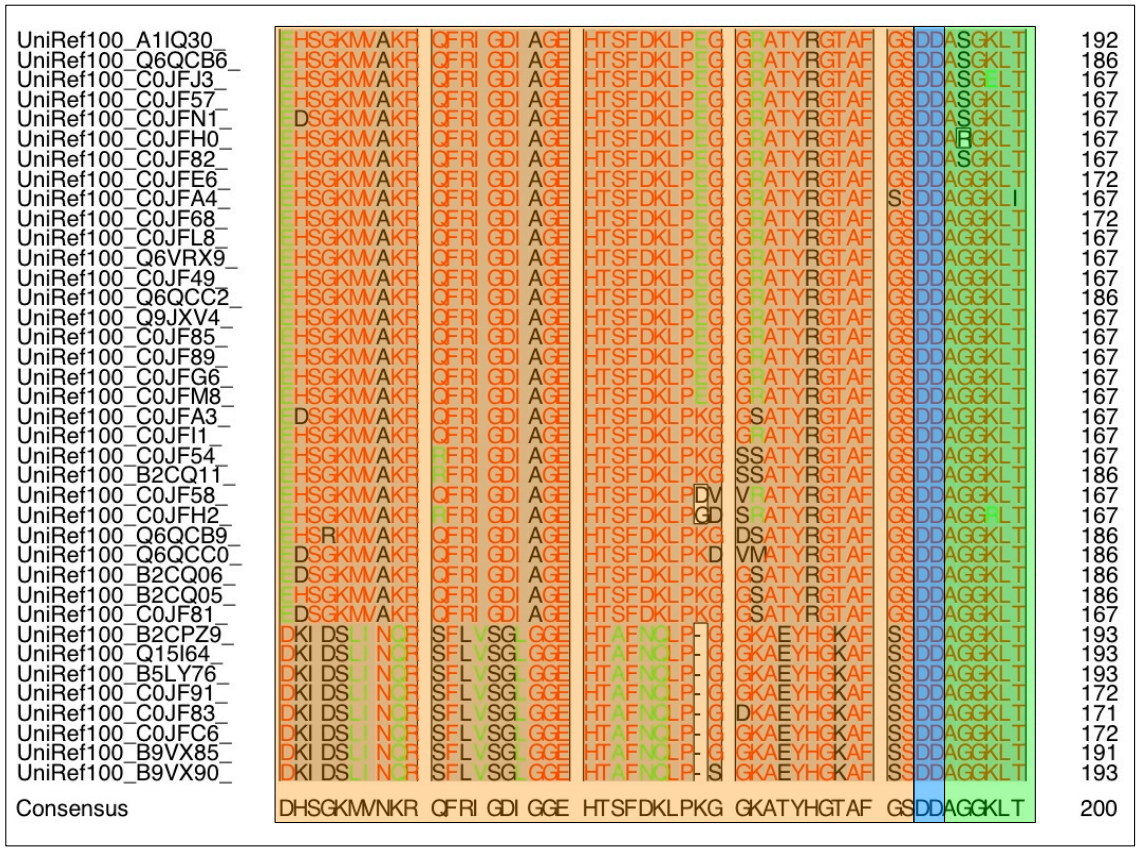


Figure C.4: A portion of the multiple sequence alignment of the 190 fHbp sequences illustrating the variable regions V_C and V_D flanked by the invariant segment 'DD'. The orange region is the continuation of the variable region V_C from Figure C.3 and variable region V_D is rendered in the lime colour. The blue region represents the invariant block 'DD' which flanks the variable regions V_C and V_D.

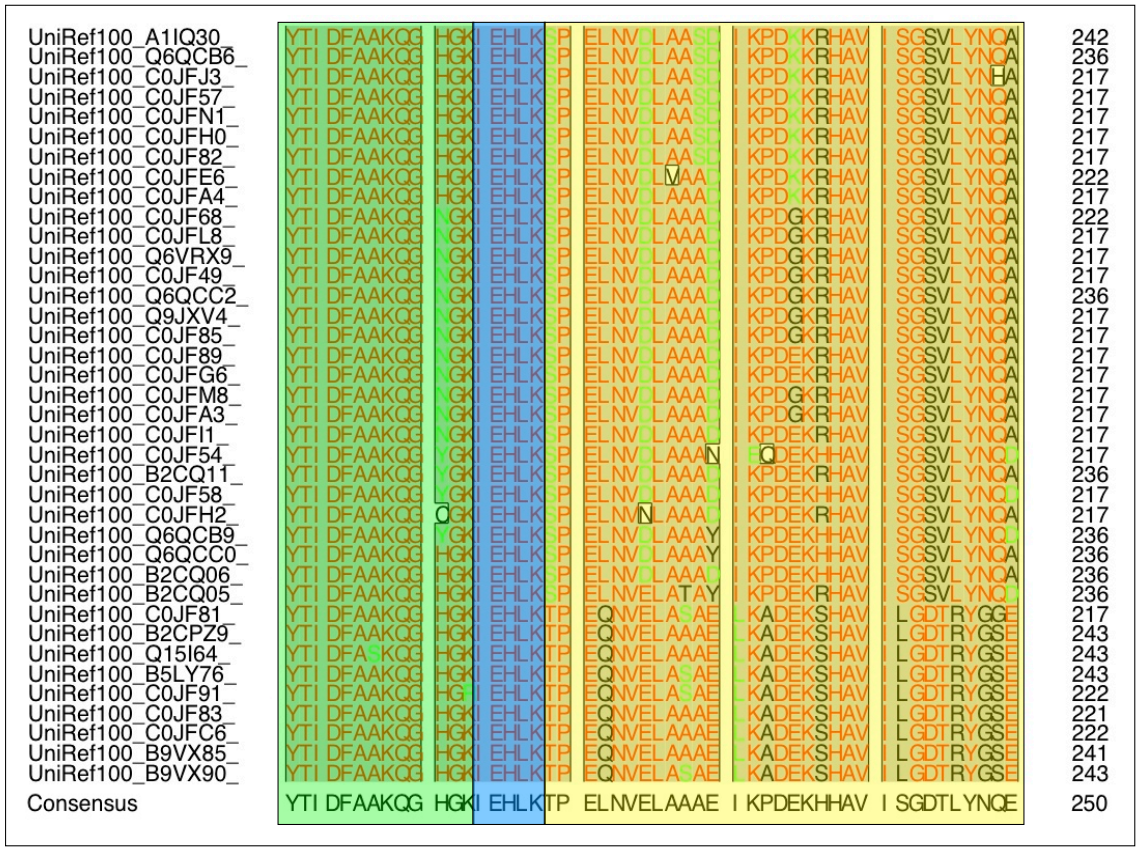


Figure C.5: A portion of the multiple sequence alignment of the 190 fHbp sequences illustrating the variable regions V_D and V_E flanked by the invariant segment 'IEHLK'. The lime coloured region is the continuation of the variable region V_D from Figure C.4 and the variable segment V_E is rendered in yellow. The blue region represents the invariant block 'IEHLK' which flanks the variable regions V_D and V_E.

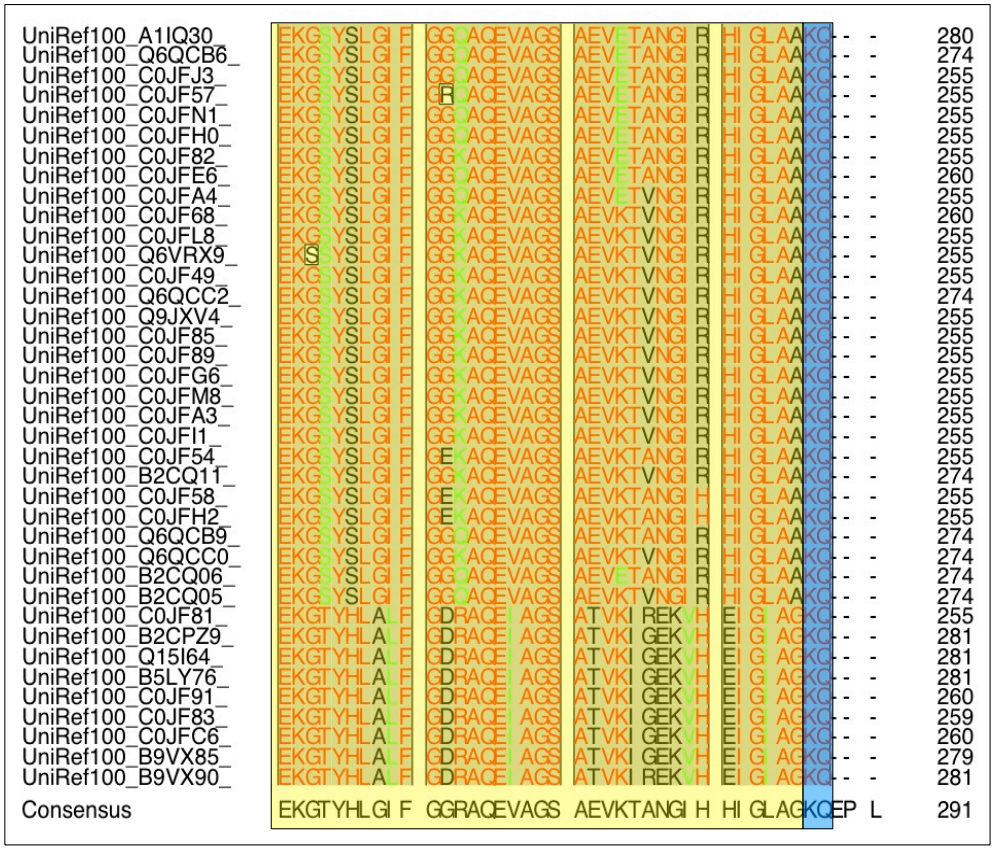


Figure C.6: A portion of the multiple sequence alignment of the 190 fHbp sequences illustrating the last variable region V_E and the last invariant segment 'KQ' rendered in blue colour. The yellow region is the continuation of the variable segment V_E from Figure C.5. The explanation of the derivation of the consensus sequence in Figure C.1 would explain the last three amino acids 'E', 'P', and 'L' even though the positions show all 'gap' symbols for all the displayed sequences. These three amino acids occurred at least for once in one of the 190 sequences and those sequences have not been illustrated in this figure, since this figure is a portion of the alignment of the 190 fHbp sequences.

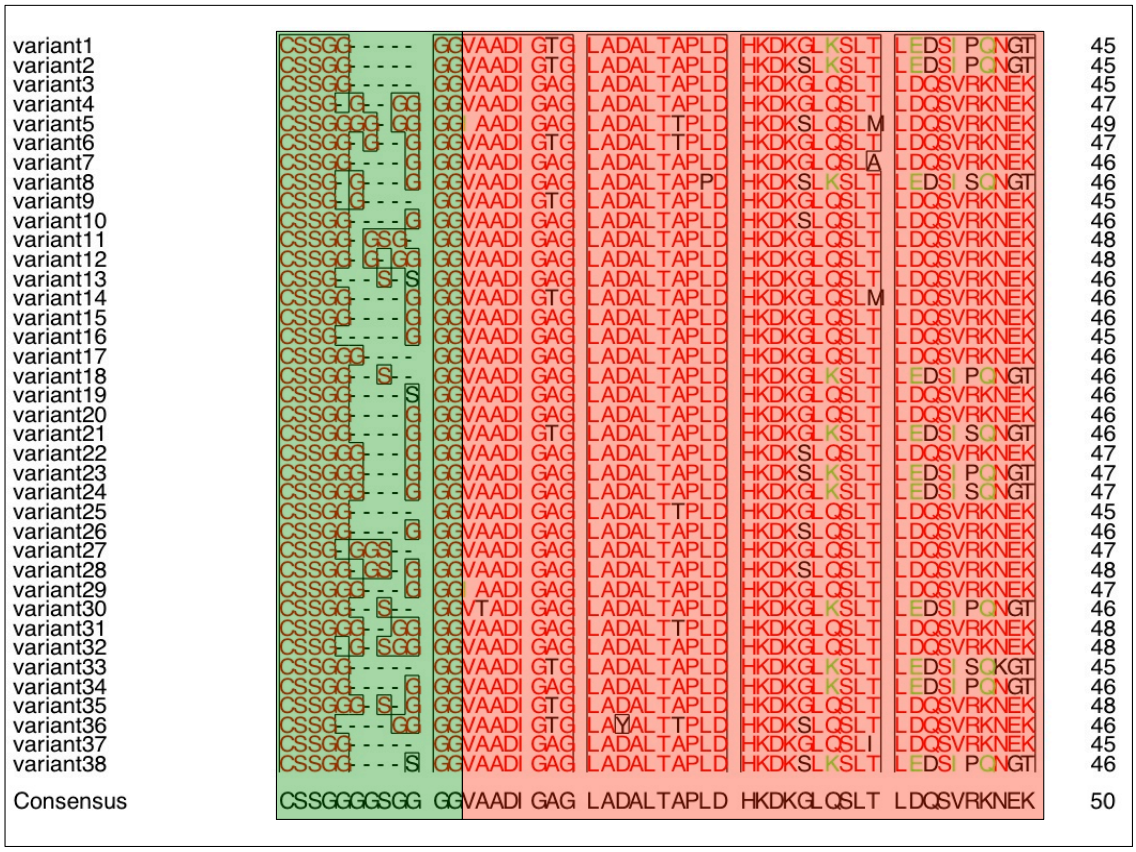


Figure C.7: A portion of the alignment of the new variants of Set A illustrating the amino-terminal repetitive element and the variable region V_A. The green region represents the amino-terminal repetitive element, which marks the beginning of the mature protein part with cysteine (the N-term region of Figure 2.2), and the red region represents the first variable region V_A (see Section 2.6 and Figure 2.2). The numbers on the right are the same as explained in Figures C.1 and C.2.

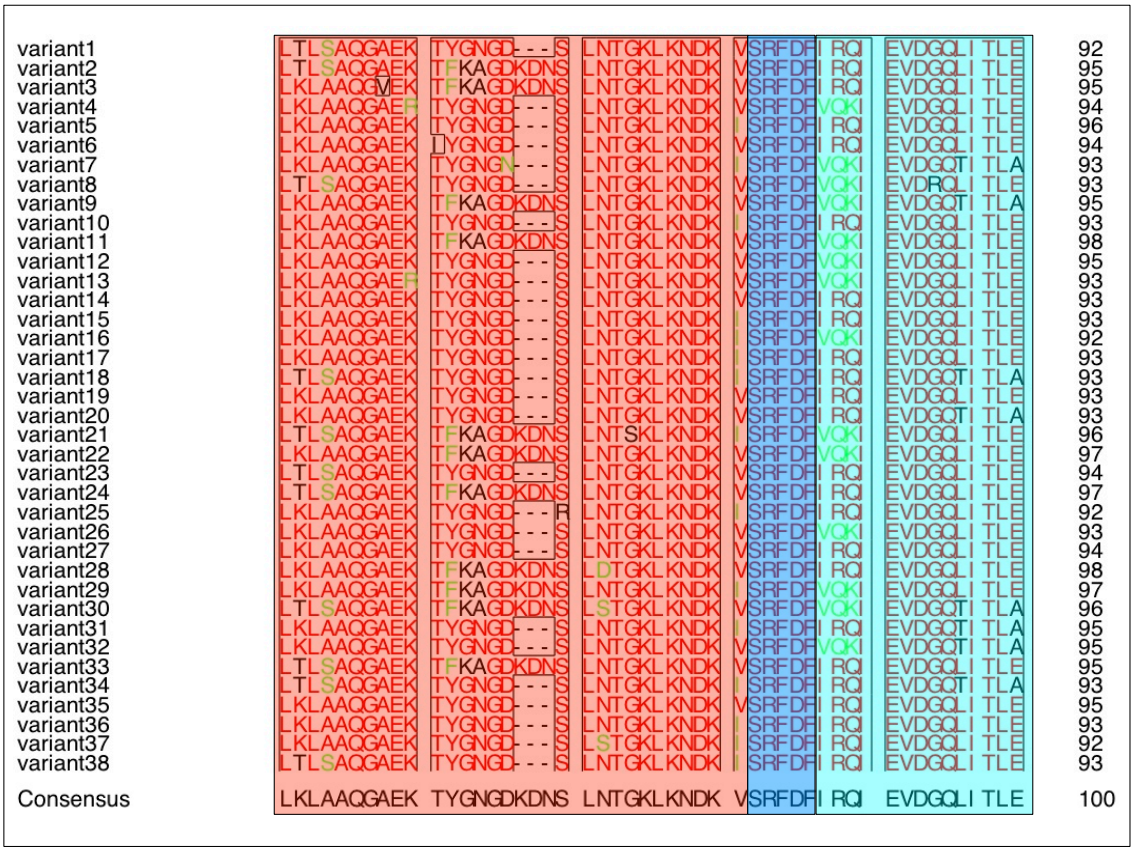


Figure C.8: A portion of the alignment of the new variants of Set A illustrating the variable regions V_A and V_B flanked by the invariant segment 'SRFDF'. The red region on the left is a continuation of V_A from Figure C.7. Region V_B is coloured in cyan and the blue region 'SRFDF' corresponds to the black vertical rectangle that separates the regions V_A and V_B in Figure 2.2.

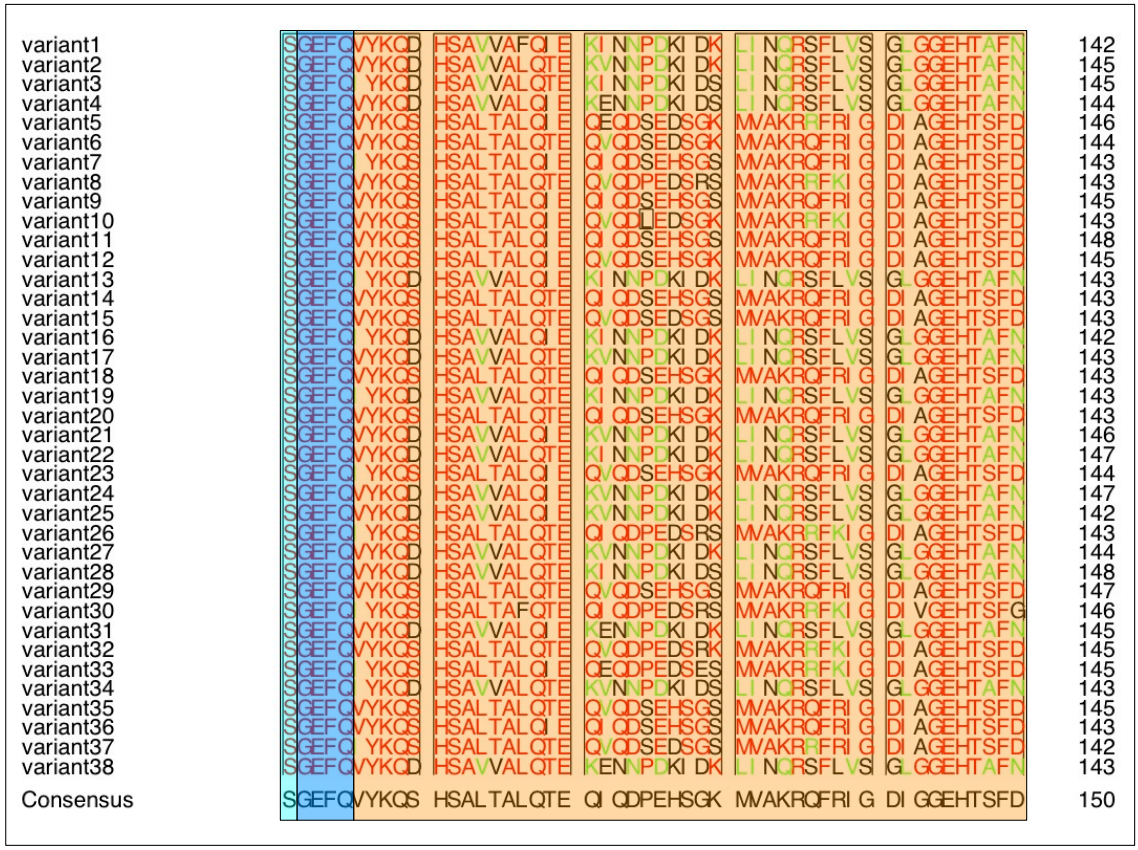


Figure C.9: A portion of the alignment of the new variants of Set A illustrating the variable regions V_B and V_C flanked by the invariant segment 'GEFQ'. The cyan coloured region is the continuation of region V_B from Figure C.8 and region V_C is coloured in orange. The blue region represents the invariant block 'GEFQ', which flanks the variable regions V_B and V_C.

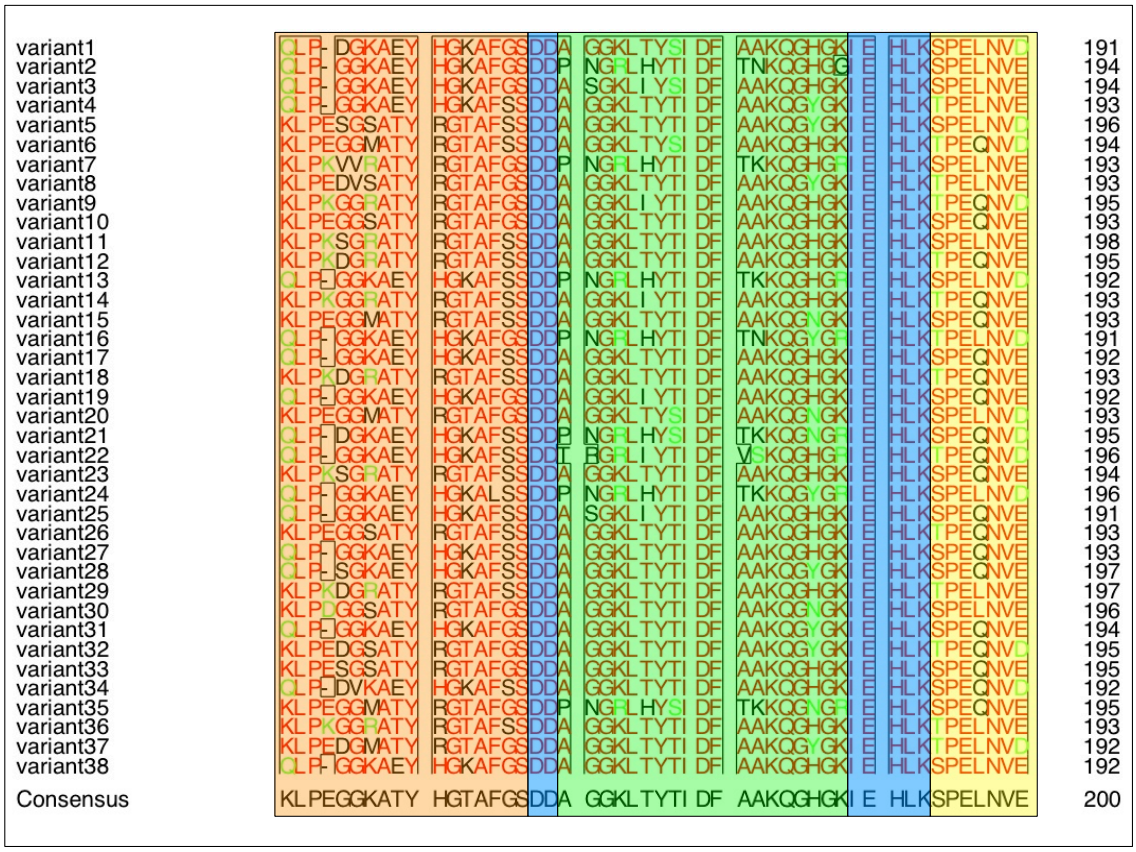


Figure C.10: A portion of the alignment of the new variants of Set A illustrating the variable regions V_C, V_D, and V_E flanked by the invariant segments. The orange coloured region is the continuation of region V_C from Figure C.9. The lime coloured region illustrates region V_D and the yellow region on the far right represents region V_E. The blue regions 'DD' and 'IEHLK' are the invariant blocks of residues that separate the modular variable regions V_C-V_D and V_D-V_E, respectively.

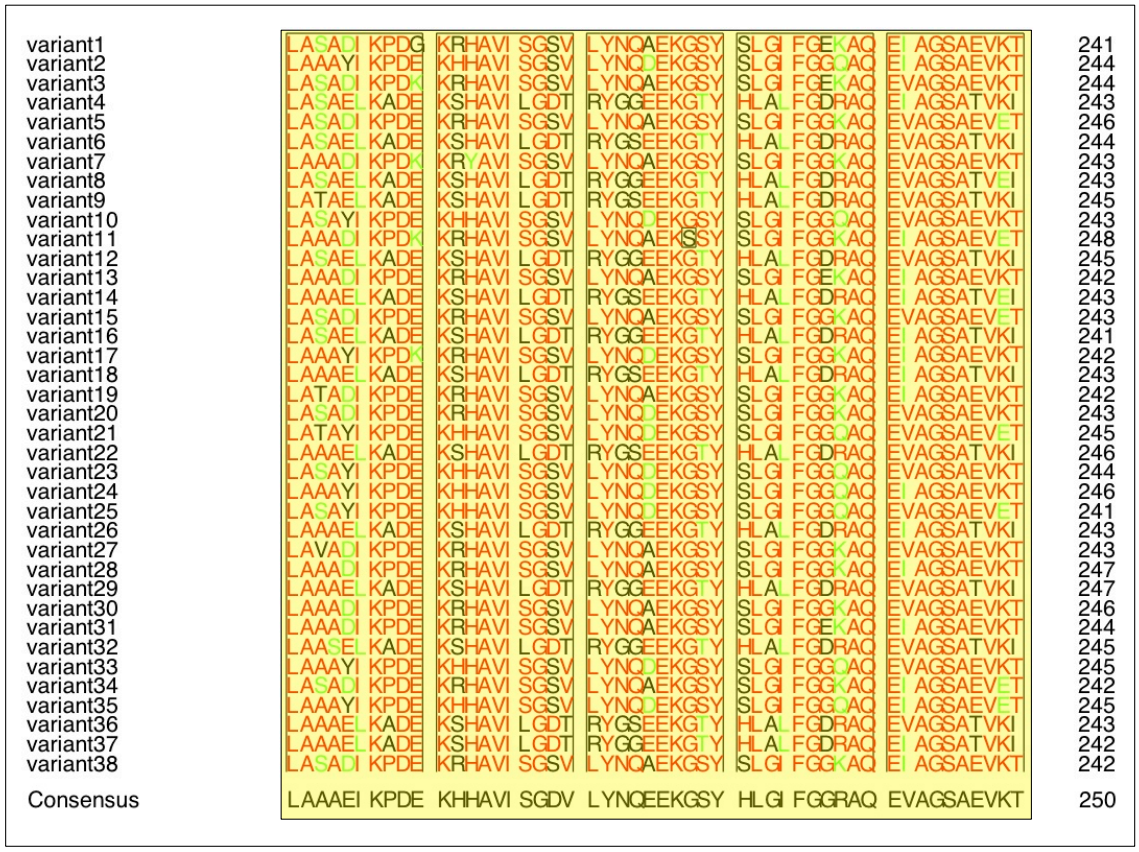


Figure C.11: A portion of the alignment of the new variants of Set A illustrating the continuation (from Figure C.10) of last variable region V_E. This whole region is highlighted in yellow.

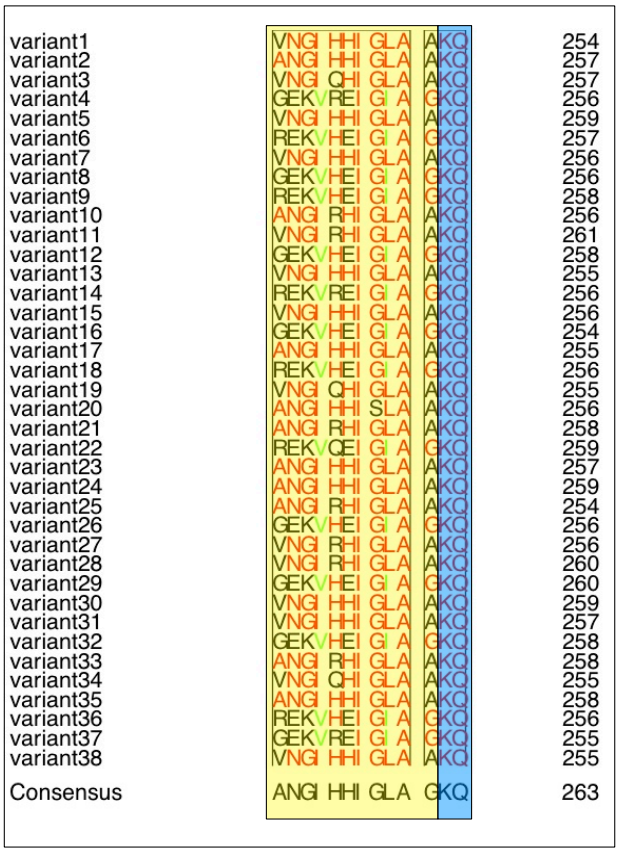


Figure C.12: A portion of the alignment of the new variants of Set A illustrating the last variable region V_E and the last invariant segment 'KQ'. The yellow region is a continuation of the region V_E from Figures C.10 and C.11. The blue region is the last invariant region, 'KQ'.

Appendix D

Amino Acid Chart

Table D.1: 20 Amino acids, their single-letter codes (SLC), and their corresponding DNA codons.

Amino Acid	**SLC**	**DNA codons**
Isloleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG